The University Of Sheffield.

Fine mapping complex disease genes with incomplete functional genomic information using shrinkage priors

Abdulaziz Ahmed Alenazi

Supervised by Dr Kevin Walters, Dr Miguel Juarez and Prof Angela Cox

Submitted for the degree of Doctor of Philosophy

The University of Sheffield, School of Mathematics and Statistics

September 2017

Abstract

Both frequentist and Bayesian approaches have successfully been used in fine mapping of complex disease genes. We thoroughly compare, by simulation, a frequentist approach (Sequential Logistic Regression (SLR)) with three other Bayesian-inspired approaches in fine mapping case-control studies: HyperLasso(HL), (PiMASS), and the Normal-Gamma (NG) prior. Our results indicate considerable variation in performance of all methods dependent on the scenario considered. The main advantage here of the Bayesian approach is that it allows inclusion of functional genomic information in the prior. We show how to include a certain form of functional information (published functional significance (FS) scores) in the NG prior whilst dealing efficiently with missing functional information.

FS scores were partitioned into four natural groups each of which was given a specific prior. We show how this approach can improve detection of the causal SNP even when it is highly correlated with several nearby SNPs. We show by using simulated case-control data that the modified NG prior can increase the true positive rates at relevant low false positive rates compared to SLR, PiMASS, HL, and the standard NG prior.

Contents

List of Figures

List of Tables

Chapter 1

Basic molecular biology and statistical genetics

This Chapter will discuss the background statistical genetics relevant to this research. The process of protein formation and genetic mutation will be discussed and statistical approaches to identifying specific mutations will be discussed briefly. Moreover, the breast cancer top hits data will be discussed at the end of this Chapter, which is used regularly in this research.

1.1 The human genome and DNA mutations

This section is intended to give a brief overview of the production of protein from deoxyribonucleic acid (DNA). The human body consists of millions of cells. The genome is all the genetic information of an organism, contained in each cell nucleus. In humans, there are two copies of each chromosome; one copy comes from the father and the other comes from the mother. There are 23 pairs of chromosomes; 22 pairs of autosomes and a pair of sex chromosomes. In terms of chromosome number, there are two types of cells: haploid and diploid cells. Haploid cells contain a single cope of each chromosome whilst diploid cells contain two copies. The sperm and egg are haploid and all other cells are diploid. The genome is made of deoxyribonucleic acid (DNA) and all genetic information is kept in this DNA.

1.1.1 The central dogma

The central dogma is that double-stranded DNA is transcribed into single stranded ribonucleic acid (RNA) which is translated into amino acids, the building blocks of proteins (see figure 1.1). We now describe the processes involved in the central dogma in detail.

Figure 1.1: The central dogma.

Deoxyribonucleic acid (DNA)

The genome is made of deoxyribonucleic acid (DNA) and DNA is made of nucleotides. The nucleotide consists of three components: a sugar, a phosphate group and a nitrogenous base, where the sugar is called deoxyribose and there are only four nitrogenous bases in DNA: adenine (A), guanine (G), cytosine(C) and thymine (T). These are carried on a ribose - phosphate backbone. DNA includes two strands that are twisted to form the double helix shape. Adenine always links with thymine by two hydrogen bonds and cytosine always links with guanine by three hydrogen bonds across the helix. Each strand has two ends, a $5'$ end and $3'$ end and these strands run opposite to each other. They are antiparallel, because one strand runs from $5' \rightarrow 3'$ whereas the second strand runs from $3' \rightarrow 5'$ see figure 1.2 for an illustration

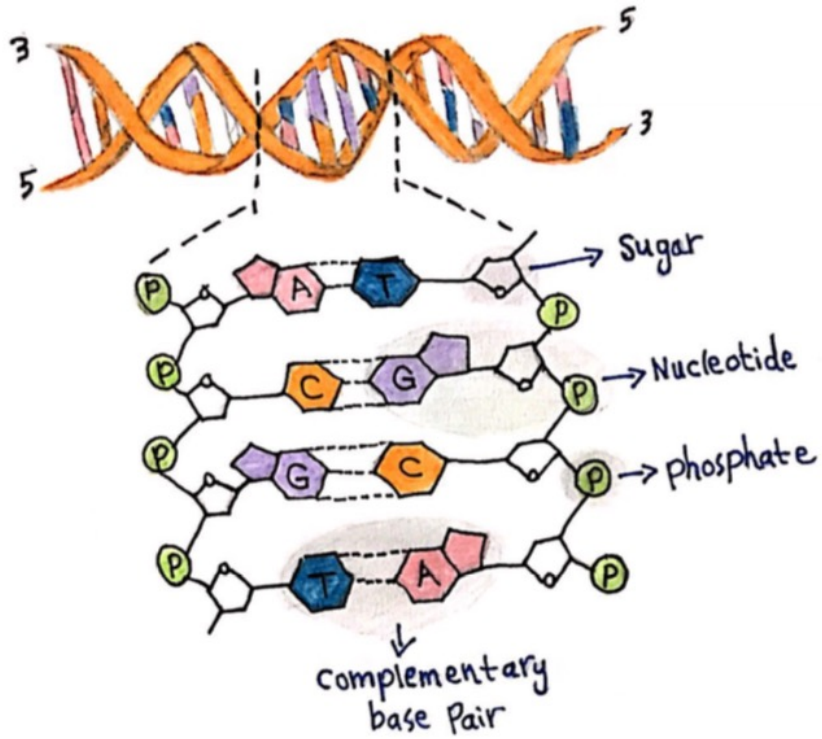


Figure 1.2: Deoxyribonucleic acid (DNA).

Deoxyribonucleic acid (DNA) replication

Here we will discuss the process of DNA replication. The two strands of parent DNA are split up into two daughter DNA strands. Initially an enzyme called helicase is used to unzip the parent DNA forming a replication fork. These two separate strands are used as templates for producing new double-stranded DNA. Another enzyme called primase begins the process by creating a primer which is a small fragment of RNA. This fragment is the starting point for creating the new DNA strands. An enzyme called DNA polymerase links to the primer to create the new strand. This polymerase is able to run only in the $5' \rightarrow 3'$ direction. The two new DNA strands created are called leading strand and lagging strand and each of them has its own process of creation. The leading strand is created in a direct continuous way where the DNA polymerase adds the complementary bases one by one from $5'$ all the way to $3'$. However, creating the lagging strand is more complicated than creating the leading strand. The reason behind this is that the lagging strand runs in the opposite direction to the leading strand. In order to create the lagging strand, the DNA polymerase makes several sequential

fragments in the $5' \rightarrow 3'$ direction (Okazaki fragments). Then an enzyme called exonuclease destroys the RNA primer from both DNA strands. Another DNA polymerase together with DNA ligase fills the resulting gaps which are left behind by the destruction of the RNA primer. Therefore, the process will result in two daughter DNA molecules. This process usually is called semi-conservative because each of the two resulting daughter DNA molecules has one strand from the parent DNA molecule (Strachan and Read, 2004).

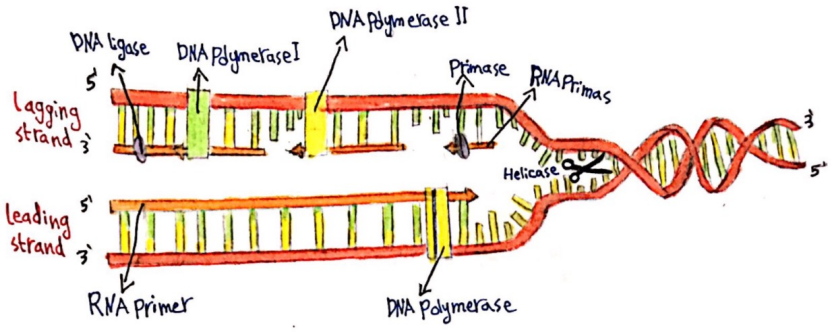


Figure 1.3: Deoxyribonucleic acid (DNA) replication.

Ribonucleic acid (RNA)

Ribonucleic acid (RNA) is a single-stranded molecule. Therefore, it has only one long sequence of nucleotides. The nucleotide includes three components: a sugar, a phosphate group and a nitrogenous base, where the sugar is called ribose and there are only four nitrogenous bases in RNA: adenine (A), guanine (G), cytosine(C) and uracil (U) (see figure 1.4).

There are three types of RNA in the cells:

1. Messenger RNA (mRNA) is a molecule that determines the polypeptides to be produced.

2. Transfer RNA (tRNA) is used to transform the amino acids from cytosol to ribosomes and also it is used to translate RNA into protein.

3. Ribosomal RNA (rRNA) is used to produce the ribosomes in the cell nucleus.

Table 1.1 summarises the major differences between DNA and RNA.

Nucleic acid	Number of strand	Nitrogenous bases	Sugar
DNA	2	Adenine (A), Guanine (G), Cytosine(C) and Thymine (T)	Deoxyribose
RNA	1	Adenine (A), Guanine (G), Cytosine(C) and Uracil (U)	Ribose

Table 1.1: The major differences between DNA and RNA

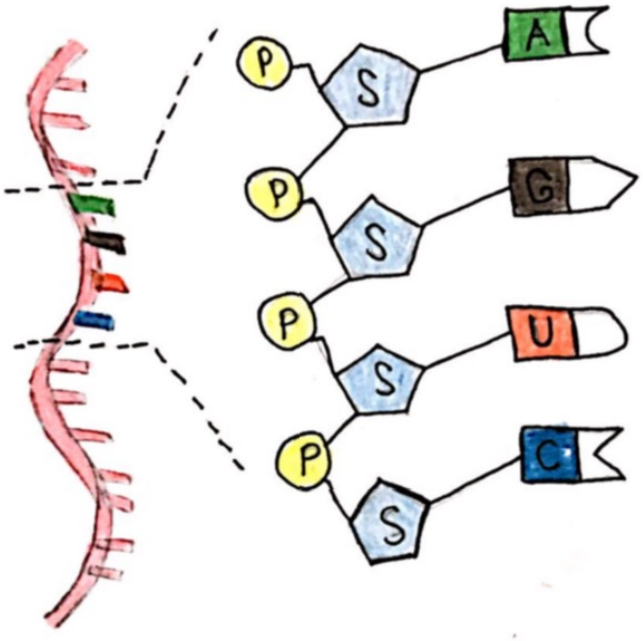


Figure 1.4: Ribonucleic acid (RNA). A represents Adenine, U refers to Uracil, C represents Cytosine, and G refers to Guanine.

Proteins

Proteins take a variety of forms such as enzymes scaffold proteins, membrane receptors, structural proteins for examples. Each protein has a particular function in each cell. The building units of proteins are the amino acids. There are only 20 amino acids that are used to produce different proteins in humans. Producing the protein is the last stage of the central dogma.

Transcription and translation

The two main operations in the central dogma are the transcription process and translation process, which transcribes the segment of DNA to mRNA and translates the resulting mRNA to protein, respectively. In this section we will discuss these processes.

Transcription is the process of transcribing a segment of DNA to RNA. To transcribe a sequence of mRNA, firstly, RNA polymerase connects to one of the DNA strands at a promoter region which is the starting region. The mRNA will be built by adding new nucleotides. Each nucleotide G on DNA will build nucleotide C on mRNA, each nucleotide C on DNA will build nucleotide G on mRNA, each nucleotide T on DNA will build nucleotide A on mRNA and each nucleotide A on DNA will build nucleotide U on mRNA instead of nucleotide T. This process will continue until the RNA polymerase reaches the stop region which is called the terminator region. The polymerase will remove from the DNA strand and the mRNA will be released. Then the mRNA is cleaned up by removal of the sequence of nucleotides that do not code for any of the 20 amino acids. These RNA sequences are called introns. Finally, the amino acid coding parts of the RNA sequences (exon) are kept and spliced together. The mRNA containing only the exons is called mature mRNA. The mature mRNA moves from the nucleus to the cytoplasm where the process of translation will occur.

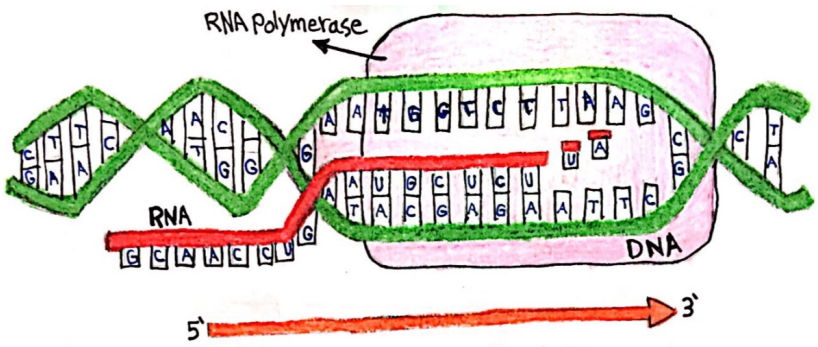


Figure 1.5: The transcription of DNA.

Translation is the process of translating the mature mRNA to proteins. The mature mRNA is a single-stranded molecule containing the exons. Within the exon three consecutive nucleotides are called a codon. Any mature mRNA has start codon that is almost always AUG and has a stop codon that can take one of the following forms: UGA, UAG or UAA stop

codons terminate translation. To translate the mature mRNA, at the beginning of the translation process a small ribosomal subunit connects to the mature mRNA at the start codon. Then tRNA comes carrying two products: the amino acid and the anticodon codon. The anticodon codon contains the three nucleotides that are the complementary nucleotides of the three nucleotides on the mature mRNA. Then the large ribosomal subunit connects to the mature mRNA to produce the active ribosome to be ready to start the process of translation at some location (say P) in the large ribosomal subunit. After that a new tRNA comes carrying a new amino acid and new three nucleotides that correspond to the next three nucleotides on the mRNA. This new tRNA links to the large ribosomal subunit at some location (say A) and peptide bonds will be built between the two amino acids curried by the tRNA at P and A. This process continues producing the polypeptides which form the proteins until the stop codon is reached (see figure 1.6). At this stage the protein releases and both the small and large ribosomal subunits disconnect. Each protein is synthesised based on the amino acid sequence encoded by the mRNA which has been transcribed from the DNA.

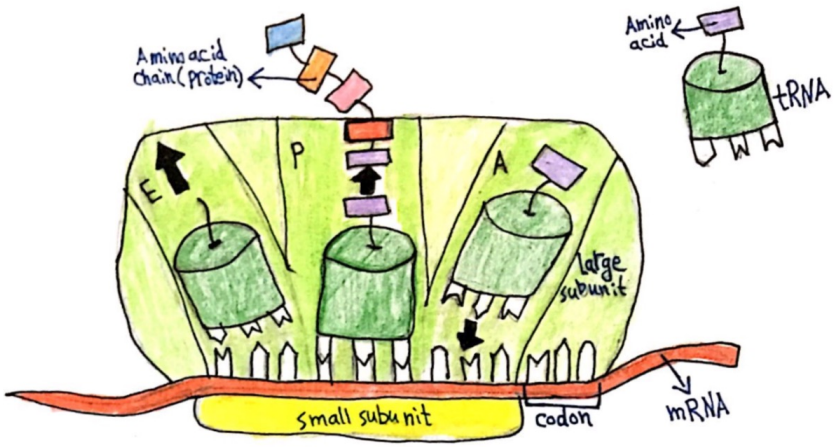


Figure 1.6: The translation of DNA.

1.1.2 DNA mutations and inheritance of mutations

In this section we will discuss inheritance of mutations including meiosis and recombination. Also, we will discuss the DNA mutations and its types.

In section 1.1, it was mentioned that everybody has two copies of each chromosome, a

copy comes from father and a copy from mother, but how are these copies inherited by the offspring? The inheritance begins in the sperm and ovum during the meiosis process. Meiosis occurs in two stages: meiosis I and meiosis II (for more details see (Strachan and Read, 2004)). In the meiosis process, the diploid progenit or haploid cell is divided into four haploid cells. These four haploid cells are not identical to those of the father and mother but contain parts of both parents. During meiosis the male and female derived chromosomes recombine (on average at one position per chromosome) to generate new chromosomes that contain genes from both parents. This process is called recombination. Recombination acts to increase genetic diversity and as a result, everybody has a unique DNA sequence that distinguishes it from others.

Recombination can occur between any two loci in the same chromosome. For instance, assume a person has two copies of a particular chromosome (say chromosome 2) and this person has genotype AB at some locus in one copy of chromosome and has genotype ab at the same locus on the second copy of chromosome. This person could produce a gamete with AB or ab. If this is the case, these are called non-recombinant. If they produce a gamete with Ab or aB, then these are called recombinant. If the two loci are physically close to each other then recombination is less likely to occur between them resulting in a non-recombinant. This tendency for nearby alleles on the same chromosome to be co-inherited is known as linkage disequilibrium (LD) (see Section 1.1.2).

During meiosis, DNA inherited from parents may also change as a result of DNA mutation. These mutations could be adding extra bases (insertions), missing bases (deletions) or rearranging bases. These three mutations are relatively rare. However, the most common DNA mutation is a Single Nucleotide Polymorphism (SNP), in which a single nucleotide is replaced with another nucleotide.

Single nucleotide polymorphism (SNP)

A Single Nucleotide Polymorphism (SNP) is a DNA sequence variation occuring in at least 1% of a population. A SNP is a single nucleotide which can be A, T, C or G in the DNA which differs between individuals. For instance, let AAGCCTA and AAGCTTA be DNA sequences from two different people; these sequences have a difference in a single nucleotide at position

5 with allele C and T respectively; this is an example of a SNP (see figure 1.7). If the first allele occurs in 70% of the population and the second allele occurs in 30% of the population, then allele T is called the minor allele and its minor allele frequency (MAF) $= 0.3$. The genotype is the two bases (alleles) carried by an individual at that SNP. An individual is homozygous at a particular SNP site if they have the same base pair for both alleles, whereas if they have a different base pair they are heterozygous. For instance, the "CC" or "TT" genotypes are homozygous whereas the "CT" genotype is heterozygous.

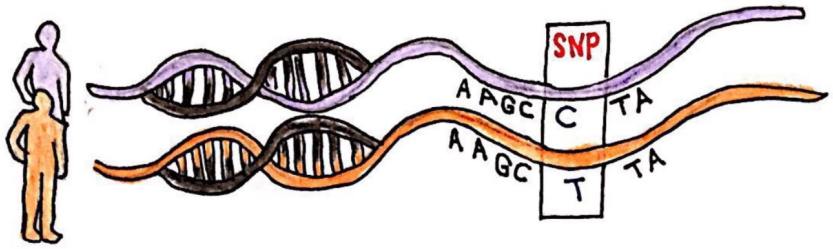


Figure 1.7: Single Nucleotide Polymorphism (SNP).

Linkage disequilibrium

A very important concept in genetic statistics is linkage disequilibrium (LD) which is used to measure allele association at two loci in the genome. LD can be measured by$D$ or D' both introduced by Lewontin (1964) and r^2 (Hill and Weir, 1994). Suppose there are two loci A and B with Allele 1 and Allele 2 at each. Let n_{ij} be the relative frequency of the haplotype having allele i at locus A and allele j at locus B (see Table 1.2). Then the allelic relative frequency at each locus is given in Table 1.3. The linkage disequilibrium (D) is defined as the difference between the haplotype relative frequency and the product of the allele relative frequencies of the two alleles at two loci (see Table 1.4). Thus, we have

$$D = n_{11} - p_1 q_1, \tag{1.1}$$

where D could be positive or negative.

Markers	Allele 1 at locus A	Allele 2 at locus A
Allele 1 at locus B	n_{11}	n_{21}
Allele 2 at locus B	n_{12}	n_{22}

Table 1.2: Haplotype relative frequency across two diallelic loci.

Markers	Allele 1
Allele 1 at locus A	$p_1 = n_{11} + n_{12}$
Allele 2 at locus A	$p_2 = n_{21} + n_{22}$
Allele 1 at locus B	$q_1 = n_{11} + n_{21}$
Allele 2 at locus B	$q_2 = n_{12} + n_{22}$

Table 1.3: The allelic relative frequency at two diallelic loci based on the haplotype frequencies.

Devlin and Risch (1995) calculated D' as follows

$$D' = \begin{cases} \frac{D}{\min(p_1 q_2, p_2 q_1)} & \text{if } D \geq 0; \\ \frac{D}{\max(-p_1 q_1, -p_2 q_2)} & \text{if } D < 0. \end{cases}$$

D' is normalising the D, hence it takes values in $[-1, 1]$. A more common measure of LD is the correlation between a pair of loci (r^2) and it is calculated as

$$r^2 = \frac{D^2}{p_1 p_2 q_1 q_2}, \tag{1.2}$$

where r^2 also takes values in $[0, 1]$. r^2 tends to be used as the LD measure because the power to detect a causal SNP at an untyped locus is directly related to the r^2 value between untyped locus and a genotyped locus. For example, suppose the power to detect a causal SNP at a genotyped locus is 80%. If the causal SNP is actually an untyped SNP in LD with the typed SNP (with a correlation of r^2) then the power to detect a causal SNP at the untyped SNP is approximately 80 times of r^2.

Markers	Allele 1 at locus A	Allele 2 at locus A	Total
Allele 1 at locus B	$n_{11} = p_1q_1 + D$	$n_{21} = p_2q_1 - D$	q_1
Allele 2 at locus B	$n_{12} = p_1q_2 - D$	$n_{22} = p_2q_2 + D$	q_2
Total	p_1	p_2	1

Table 1.4: Relationship between haplotype relative frequency, allelic relative frequency, and the deviation (D).

1.2 Genetic association studies

Population-based genetic association studies attempt to establish statistical associations between genetic sequence information (i.e. variation in the nucleotide sequences of the genomes) of unrelated individuals in a population and the measurable traits of those individuals, in particular the presence or progression of complex diseases, which are thought to be associated with multiple genetic polymorphisms. There are four main types of population genetic association studies: Candidate Polymorphism Studies, Candidate Gene Studies, Fine Mapping Studies and Genome-Wide Association Studies (GWAS) (Foulkes, 2009). In this section, we will discuss briefly the four types of genetic association studies. Moreover, we will discuss the basic of statistical genetics and the types of model used.

An important concept of statistical genetics is penetrance which represents the proportion of carriers of a variant (SNP or nutation) who exhibit the trait associated with that variant. There are many statistical approaches having high power to identify variants with high penetrance , but these do not necessarily have high power to identify variants with lower penetrance , for example linkage studies. However, there are some approaches such as Genome-Wide Association Studies (GWAS) and fine-mapping studies that can identify variants having lower penetrance.

1.2.1 Candidate polymorphism studies

These investigate association between a particular SNP and the disease trait. These are carried out when there is already scientific evidence to suggest that the SNP is functional. The purpose of the study is to determine if the SNP directly influences the disease trait (Foulkes, 2009).

1.2.2 Candidate gene studies

These investigate association between multiple SNPs in the same gene and the disease trait. These SNPs are not necessarily functional or causative of the disease trait. The true disease-causing locus may be unknown. The SNPs act as 'markers' of the true locus and are chosen based on linkage disequilibrium (LD) between the SNPs and the true disease-causing locus, which is caused by a low probability of recombination at points between the disease-causing locus and the SNP loci. The SNPs are, therefore, presumed to be physically proximate on the genome to the disease-causing locus (Foulkes, 2009).

1.2.3 Genome-wide association studies (GWAS)

These investigate association between SNPs and the disease trait, but without the prior information present in the candidate studies. The aim is to cover large numbers of SNPs across the entire genome. The pre-processing of this data requires a large amount of processing power and specialised software which caters to the high-dimensional data (Foulkes, 2009). In order to identify putative causal SNPs, many researchers use 5×10^{-8} as a p-value threshold in their standard hypothesis test, see for example (Easton et al., 2007; Fachal and Dunning, 2015). GWAS studies typically contain a million SNPs. Therefore, the 5×10^{-8} comes from a Bonferroni correction for multiple testing where the number of tests is one million $\left(\frac{0.05}{10^6} = 5 \times 10^{-8}\right)$.

1.2.4 Fine mapping studies

In Genome-Wide Association Studies (GWAS), many regions are identified as having potentially causal SNPs based on the p-values obtained. It is difficult to determine a particular causal SNP in that region. Fine mapping studies look at the specific issue of narrowing down the set of potentially interesting SNPs. This could either be by creating a ranked list of SNPs and selecting a given $\%$ for further study, or by a formal decision-theoretic process using Bayes factors and posterior odds (much less common) or by using likelihood ratios (Kichaev et al., 2014; Udler et al., 2010).

1.2.5 Statistical genetics

In this section we will talk about the elementary aspects of statistical genetics that include odds, odds ratios, statistical model, linkage disequilibrium, univariate approaches, p-values and Bayes factors and their limitations.

Odds and odds ratio

Odds in a genetic sense is the probability of being a case given a specific genotype divided by the probability of being a control given that genotype. The odds ratio (OR) is a tool to measure the association between an exposure (genotype in this case) and an outcome (case-control status). The OR refers to the ratio of the odds of being a case conditional on having a specific genotype to the odds of being a case conditional on another specific genotype (Szklo and Nieto, 2012).

Statistical model

A genetic model is defined by Foulkes (2009) as "*the biological interaction between alleles on homologous chromosomes*". We will talk about three genetic models: the additive, the dominant, and the recessive. Let us take a simple example for the genetic models. Suppose the simple situation where the alleles are "A" and "G" at a particular SNP, where "A" is the major allele while "G" is the minor allele. In addition, suppose that we are looking for a binary output, y, which takes two values, either 0 or 1 (control and case respectively). An additive model means that the effect of having a single copy of the "G" allele is to increase log odds of being a case ($y = 1$) by an amount equal to β, and having two copies of the "G" allele will increase log odds of being a case ($y = 1$) by 2β. Let $I(x_{i,k} = G)$ be an indicator for whether the allele (x) on the kth homolog $(k = 1, 2)$ is equal to "G" for individual i, then an additive genetic model for this SNP is expressed as

$$P(y_i = 1) = \frac{\exp\{\alpha + \beta[I(x_{i,1} = G) + I(x_{i,2} = G)]\}}{1 + \exp\{\alpha + \beta[I(x_{i,1} = G) + I(x_{i,2} = G)]\}}. \tag{1.3}$$

However, a dominant genetic model supposes that having at least one copy of the "G" allele will result in an increase of β in the log odds of being case and is given by the following expression

$$P(y_i = 1) = \frac{\exp\{\alpha + \beta[I(x_{i,1} = G \text{ or } x_{i,2} = G)]\}}{1 + \exp\{\alpha + \beta[I(x_{i,1} = G \text{ or } x_{i,2} = G)]\}}. \tag{1.4}$$

The third type of model considered is a recessive model that supposes that both homologs must include the rare allele in order for the effect to be present. The effect is that for individuals with 2 copies of the rare allele the log odds of being a case ($y = 1$) increases by β relative those not carrying 2 copies of the rare allele. This model is expressed as

$$P(y_i = 1) = \frac{\exp\{\alpha + \beta[I(x_{i,1} = G \text{ and } x_{i,2} = G)]\}}{1 + \exp\{\alpha + \beta[I(x_{i,1} = G \text{ and } x_{i,2} = G)]\}}. \tag{1.5}$$

In this project we will consider only the additive model as this is by far the most common form of genetic model identified in fine-mapping studies to date.

Univariate approaches and its limitations

In fine mapping studies researchers often apply univariate approaches such as p-values which are the most popular summary measure for inference in GWAS (Balding, 2006). Univariate logistic regression is most commonly used in fine-mapping studies to calculate p-values. In a Bayesian framework univariate Bayes factors (BFs) are an alternative to p-values which were studied by Wakefield (2009) in context of GWAS. He defined (BF) as the ratio of the probability density of the alternative hypothesis over null hypothesis and it is given as follows

$$\text{BF} = \frac{f\left(data \mid H_1\right)}{f\left(data \mid H_0\right)}. \tag{1.6}$$

Moreover, he showed BFs could be used alongside posterior odds in a formal decision-theoretic framework. However, these univariate methods ignore the relationship between SNPs. Therefore, they lose some information by not considering the joint distributions of SNPs when there is more than one causal SNP in the region. Moreover, these methods are likely to be less powerful than the multiple SNP approaches such as stepwise logistic regression, PiMASS, Hyper lasso and Normal-Gamma prior considered in this thesis.

1.3 Published top hits from large breast cancer genome wide association studies

In this thesis we take a Bayesian approach to fine-mapping and so we need to put a prior on the effect size (β). Wakefield (2008) suggested elicitation from an expert statistical geneticist. To avoid the subjective nature of this procedure we choose to use the top hits identified in several large-scale breast cancer GWAS in addition to early identified BRCA mutations. Although most of the variants discovered (at p-values $< 5 \times 10^{-8}$) have not yet been validated they are strong candidates for causal SNPs, and assuming they do tag actual causal SNPs, the true effect sizes are likely to be close to those of the top hits identified.

Several genome wide association studies were used to identify SNPs associated with breast cancer. The first five potential causal SNPs were identified in 2007 (Easton et al., 2007) followed by the identification of another 68 potential causal SNPs using data from both European and Asian ancestry reviewed in Fachal and Dunning (2015). According to Fachal and Dunning (2015), most of the observed odds ratios for these causal SNPs were between 1.05 and 1.26. Moreover, they reported that the odds ratio of approximately 1.05 is the smallest GWAS significant odds ratios discovered, at a p-value of $< 5 \times 10^{-8}$ threshold.

In a further study, Michailidou et al. (2015) applied a meta-analysis of 11 genome wide association studies (GWAS) including $15,748$ cases and $18,084$ breast cancer controls along with $46,785$ cases and $42,892$ controls from 41 studies, which were genotyped on a more than $200,000$ marker custom array (iCOGs). Additionally, they used 1000 Genomes Project (Altshuler et al., 2010) as a reference data set to impute more than 11 million SNPs, having set $r^2 > 0.3$ and minor allele frequency (MAF)> 0.005. They applied GWAS to $120,000$ individuals of European ancestry. The threshold they selected to identify the association between breast cancer and SNPs was the usual p-value $< 5 \times 10^{-8}$ threshold. In their study, a 15 further SNPs were discovered as candidate causal SNPs for breast cancer. Furthermore, 65 other SNPs were identified as a potential causal SNPs associated with the breast cancer for more information see (Stacey et al., 2007, 2008; Zheng et al., 2009; Ahmed et al., 2009; Thomas et al., 2009; Turnbull et al., 2010; Fletcher et al., 2011; Ghoussaini et al., 2012; Siddiq

et al., 2012; Michailidou et al., 2013a; Cai et al., 2014; Milne et al., 2014).

In the first two studies (Fachal and Dunning, 2015; Michailidou et al., 2015), 83 SNPs were identified as potential causal SNPs for breast cancer and the causal SNPs discovered explain about 14% of the familial risk of breast cancer. However, Michailidou et al. (2013b) used Quantile-Quantile (Q-Q) plots to estimate that there may be up to 1000 additional common causal SNPs having small effect sizes that are yet-to-be discovered. The vast majority of these yet-to-be discovered variants have odds ratios between 1.02 and 1.05. These undiscovered variants have been estimated to explain an additional 14% of the familial risk of breast cancer.

Henceforth we will refer to 148 potential causal SNPs so far as the "breast cancer top hits" data or BCTH data. We considered the 68 potential causal SNPs identified by Fachal and Dunning (2015), the 15 potential causal SNPs identified by Michailidou et al. (2015) and 65 potential causal SNPs identified by others see (Stacey et al., 2007, 2008; Zheng et al., 2009; Ahmed et al., 2009; Thomas et al., 2009; Turnbull et al., 2010; Fletcher et al., 2011; Ghoussaini et al., 2012; Siddiq et al., 2012; Michailidou et al., 2013a; Cai et al., 2014; Milne et al., 2014).

When deriving effect size priors we will use both the observed estimates of the log odds ratios of the top hits and the fact that there are likely to be a large number of additional potential causal SNPs in breast cancer.

Chapter 2

Multivariate statistical models

In the previous Chapter we discussed basic molecular biology including the central dogma, DNA mutations and inheritance of mutations, but also discussed the statistical genetics including genetic association studies and statistical model. Moreover, the published top hits from large breast cancer genome wide association studies were discussed.

Variable selection is an important objective for researchers nowadays. In frequentist statistics, stepwise regression is often applied in various ways, forward, backward or combined. Bayesian approaches to variable selection are also common. In fine-mapping, these Bayesian approaches include Hyper lasso and PiMASS. They employ shrinkage priors to induce model sparsity. We also consider in detail an alternative shrinkage prior (the Normal-Gamma prior) that has never been considered in fine-mapping association studies.

In this Chapter we will give a literature review for frequentist approaches (stepwise logistic regression) and Bayesian approaches (Hyper lasso and PiMASS) used in fine-mapping studies and genome wide association studies.

In this Thesis, we will use lower-case for scalars, lower-case bold for vectors, and upper-case bold for matrices.

2.1 Multiple logistic regression

In this Section, we demonstrate how to calculate and interpret the coefficients of logistic regression. It is a particular type of generalised linear models (GLM) (Nelder and Baker,

1972) and one of the most commonly applied techniques with data with a binary response variable.

We have case-control data so the response variable is a binary variable which takes the value 1 if it is a 'case' and 0 if it is a 'control'. We are interested in finding the relationship between the binary response variable $\boldsymbol{y}$ (case-control status) and p categorical independent variables $\boldsymbol{x}_j$, where $i = 1, \ldots, n$ and $j = 1, \ldots, p$, which in our situation are the SNPs. n denotes the sample size and p the number of SNPs. We only consider additive models on the log risk scale so x_{ij}, takes the values 0, 1 and 2 according to the number of rare alleles.

Logistic regression focuses on $p_i = \Pr(y_i)$. Since $y_i \sim Ber(p_i)$ and are assumed to be independent the likelihood can be expressed as

$$\prod_{i=1}^{n} p_i^{y_i}(1-p_i)^{1-y_i}, \tag{2.1}$$

where $p_i \in [0, 1]$. The question is how one can model p_i given some data, denoted by $\boldsymbol{X}$. The matrix $\boldsymbol{X}$ is called design matrix and the first column is a vector of 1s. To make the model linear a logit transformation $\log\frac{p_i}{1-p_i}$ can be used. This has the advantage over other transformations that it confers a meaningful interpretation to the parameters of the statistical model.

The logistic model is then expressed as

$$\text{logit}\left(p_i\left(\boldsymbol{x}_i;\boldsymbol{\beta}\right)\right) = \log\frac{p_i\left(\boldsymbol{x}_i;\boldsymbol{\beta}\right)}{1-p_i\left(\boldsymbol{x}_i;\boldsymbol{\beta}\right)} = \boldsymbol{x}_i^T\boldsymbol{\beta}, \tag{2.2}$$

where $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^T$ is the vector of beta coefficients and $\boldsymbol{x}_i$ is a column vector of explanatory variables starting with 1 for the intercept term. Rearranging for p_i, this yields

$$p_i\left(\boldsymbol{x}_i;\boldsymbol{\beta}\right) = \frac{\exp\left(\boldsymbol{x}_i^T\boldsymbol{\beta}\right)}{1+\exp\left(\boldsymbol{x}_i^T\boldsymbol{\beta}\right)} = \frac{1}{1+\exp\left\{-\left(\boldsymbol{x}_i^T\boldsymbol{\beta}\right)\right\}}. \tag{2.3}$$

When we fit logistic regression models we have to know how to interpret the coefficients obtained from the model. In linear regression, the β_j shows the expected change in the response variable $\boldsymbol{y}$ as the independent variable $\boldsymbol{x}_j$ increases by one unit, while keeping all the other variables at the same unit. On the other hand, in the logistic regression, the expected

change in logit (p) is equal to β_j as we increase the independent variable $\boldsymbol{x}_j$ by one unit so β_j is the increase in log odds of disease for each additional copy of the risk allele at SNP$_j$. We focus on logistic regression with multiple SNPs so the question is how can one estimate the vector of log odds $\boldsymbol{\beta}$ and the variance-covariance matrix in this case. These can be calculated using the Newton-Raphson method (Casella and Berger, 2002). It uses the current value $\boldsymbol{\beta^{(i)}}$ for $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^T$ and calculates the updated value via

$$\boldsymbol{\beta^{(i+1)}} = \boldsymbol{\beta^{(i)}} + \left[-\ell''\left(\boldsymbol{\beta^{(i)}}\right)\right]^{-1} \ell'\left(\boldsymbol{\beta^{(i)}}\right). \tag{2.4}$$

The log likelihood of the parameters in a logistic regression model is given by

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{n} \log p_i\left((\boldsymbol{x}_i; \boldsymbol{\beta})\right)^{y_i} \left(1 - p_i\left((\boldsymbol{x}_i; \boldsymbol{\beta})\right)\right)^{1-y_i}. \tag{2.5}$$

The first derivative of $\ell(\boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}$, in matrix form is given by

$$\ell'(\boldsymbol{\beta}) = \boldsymbol{X}^T\left(\boldsymbol{y} - \boldsymbol{p}\left(\boldsymbol{x}; \boldsymbol{\beta}\right)\right), \tag{2.6}$$

where $\boldsymbol{p}\left(\boldsymbol{x}; \boldsymbol{\beta}\right)$ represents $E\left(\boldsymbol{Y}\right)$. The second derivative of of $\ell(\boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}$ can be expressed in matrix form as

$$\ell''(\boldsymbol{\beta}) = -\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X} \tag{2.7}$$

where $\boldsymbol{W}$ is the diagonal matrix with diagonal elements $p_i\left(\boldsymbol{x}_i; \boldsymbol{\beta}\right)\left(1 - p_i\left(\boldsymbol{x}_i; \boldsymbol{\beta}\right)\right)$. Equation (2.4) is iterated until the $\boldsymbol{\beta}$ vector of parameter estimates has converged. If this is the case the estimates are the maximum likelihood estimates. The variance covariance matrix of $\boldsymbol{\beta}$ is given by the inverse of the Fisher information matrix or the negative of the inverse of (2.7). For more details see (Casella and Berger, 2002).

2.2 Stepwise logistic regression

Generally, applying multiple logistic regression to all SNPs directly in frequentist fine-mapping analysis is not recommended, because of multicollinearity which could lead to instability in

parameter estimates. Therefore, stepwise regression is used to select an appropriate statistical model for the given data. There are three kinds of stepwise regression: forward, backward and combined. Here, we will focus only on forward stepwise regression because this method is generally used in fine-mapping study. The algorithm of the forward stepwise regression in logistic regression is as follows

1. Start with minimal model which just contains the intercept term.
2. Apply univariate regression for each variable so p regressions are run.
3. Keep the variable having the minimum Akaike information criterion (AIC), if it is less than that of the minimal model.
4. Calculate the minimum AIC, if a variable was added at 3, and consider adding each of the remaining $p-1$ variables separately.
5. Keep the variable added in 4 that has the smallest AIC if it was less than that in 3.
6. Repeat the process until no more variables are added.

The Akaike information criterion is given by

$$\text{AIC} = 2p - 2\ln\left(L\left(\hat{\beta}; y, x\right)\right), \tag{2.8}$$

where p is the number of parameters and $L\left(\hat{\beta}; y, x\right)$ is the maximum value of the likelihood function.

In genome wide association studies (GWAS) and fine mapping studies researchers such as French et al. (2013) and Glubb et al. (2015) applied a specific method to chose the appropriate model for the data via the forward stepwise regression. Because of the potentially large number of SNPs that could be added in the models they filter the SNPs before applying stepwise regression. The algorithm they applied is as follows

1. Perform univariate logistic regression for each of the p SNPs.
2. Calculate the p-value of those p SNPs.

3. Keep the SNPs with p-values less than the given threshold and remove the rest.

4. Apply forward stepwise logistic regression on those SNPs retained in 3.

Note that the p-value threshold is specified based on the family wise error, which controls the probability of rejecting the null at least once when multiple hypotheses are tested.

This strategy was applied when analysing our case-control data using logistic regression. In order to compare this method to other Bayesian methods via a ROC curve, we calculated the true positive rate and false positive rate for different p-value thresholds used to initially filter the SNPs. This gives a set of points that can be plotted in ROC space. We followed the same strategy as French et al. (2013) and Glubb et al. (2015). We counted the number of causal SNPs that were chosen in the final model and the number of chosen non-causal SNPs. We calculated the false positive rate (FPR) and the true positive rate (TPR) as follows

$$\text{FPR} = \frac{\text{Number of selected non-causal SNPs}}{\text{Total number of true non-causal SNPs}} \tag{2.9}$$

$$\text{TPR} = \frac{\text{Number of selected causal SNPs}}{\text{Total number of true causal SNPs}}. \tag{2.10}$$

We used these values to draw points in the ROC space. For example, suppose the case control data has 100 SNPs, 5 of which are causal SNPs, and the p-value threshold is 10^{-4}. First filter the SNPs by applying univariate logistic regressions. Assume 80 SNPs, 4 of which are causal, were kept. After applying the forward stepwise logistic regression, the final model has 10 SNPs 2 of which are causal. Thus, the FPR and TPR are given as follows

$$\begin{aligned} \text{FPR} &= \frac{8}{95} = 0.08 \\ \text{TPR} &= \frac{2}{5} = 0.4. \end{aligned}$$

This (FPR,TPR) co-ordinate can then be plotted in ROC space. We vary the p-value threshold to end up with a set of (FPR, TPR) co-ordinates that we can plot.

2.3 Hyper lasso

Tibshirani (1996) proposed the least absolute shrinkage and selection operator (LASSO) which minimises the residual sum of squares subject to the sum of the absolute value of the coefficients being less than a constant. Its aim is to shrink some coefficients to exactly zero. He used a double exponential distribution (DE) as penalising function or equivalently as a prior from a Bayesian perspective (see section 2.3.2). Hoggart et al. (2008) developed the Hyper lasso with a more flexible prior distribution, the normal exponential gamma distribution (NEG).

Hoggart et al. (2008) exploited new developments in stochastic search methods to show the feasibility of simultaneously analysing, within a few hours on a desktop computer, all SNPs in a genome-wide study to identify the subset which best predicts disease outcome and explains case-control status with a specified family-wise error rate, and to demonstrate that this produces better SNP identification than single-SNP studies using the Armitage Trend Test (ATT), because it takes into the account the joint effects of SNPs. Additive, dominant and recessive contributions to disease risk can all be implemented.

In this section, we will give an over view of the Hyper lasso (HL). Moreover, we will demonstrate how one can apply the Hyper lasso software and how to specify the parameters. Advantages and disadvantages of the Hyper lasso will be discussed.

2.3.1 Shrinkage priors

The concept of variable selection is now used widely because of the abundance of data containing a huge number of variables such as genetic data. Therefore, reducing the number of variables becomes increasingly important. Bayesian approaches are commonly used to overcome this problem. Mixture priors (e.g spike and slab) are used to separate the signal and noise in many Bayesian variable selection approaches. An alternative to mixture priors are absolutely continuous priors that shrink non significant variables towards zero but apply tittle shrinkage to large effect sizes (Bhattacharya et al., 2015, 2012; Griffin and Brown, 2010). Here we will discuss a common shrinkage prior.

Selecting independent shrinkage priors having a density that is clearly peaked at zero is a

common selection for a regression coefficient (β). A common shrinkage prior is the double exponential distribution (DE), which has only one parameter. The prior distribution of β can be written as a scale mixture of normal distributions as follows

$$\begin{aligned} DE(\beta|\xi) &= \int_0^\infty N(\beta|0,\sigma^2)\ Ga(\sigma^2|1,\frac{\xi^2}{2})d\sigma^2 \\ &= \frac{\xi}{2}\text{exp}\{-\xi|\beta|\}, \end{aligned} \tag{2.11}$$

where $N\left(. \mid a, b\right)$ is the usual Gaussian probability density with mean a and variance b and $Ga\left(a, b\right)$ is a gamma distribution with shape a and scale b with probability density given by

$$Ga\left(a, b\right) = \frac{1}{\Gamma\left(a\right) b^a} x^{a-1} e^{-\frac{x}{b}}. \tag{2.12}$$

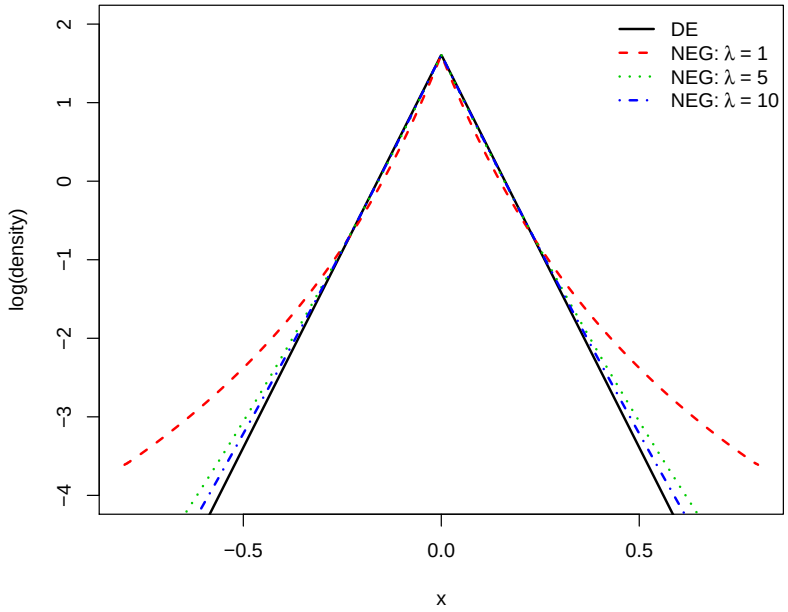


Figure 2.1: The log density of the double exponential distribution (solid line) and the log density of normal exponential gamma distribution for $\lambda = 1$ (dashed line), $\lambda = 5$ (dot line) and $\lambda = 10$ (dashed dot line). This plot is reproduced using the same values used in Hoggart et al. (2008).

Moreover, the normal-exponential-gamma distribution (NEG) (Hoggart et al., 2008) can

be used as a shrinkage prior, which is a generalisation of the DE. It is a two parameter distribution (λ, γ) and it can be written as follows

$$\begin{aligned} NEG(\beta|\lambda,\gamma) &= \int_0^\infty \int_0^\infty N(\beta|0,\sigma^2) Ga(\sigma^2|1,\Psi) Ga(\Psi|\lambda,\gamma^2) d\sigma^2 d\Psi \\ &= \kappa \ \exp(\frac{\beta^2}{4\gamma^2}) \ D_{-2\lambda-1(\frac{|\beta|}{\gamma})}, \end{aligned} \tag{2.13}$$

where D is the parabolic cylinder function (Abramowitz and Stegun, 1964) and is given by

$$D_p(z) = 2^{\frac{p}{2}} e^{-\frac{z^2}{4}} \left\{ \frac{\sqrt{\pi}}{\Gamma\left(\frac{1-p}{2}\right)} \Omega\Big(-\frac{p}{2}, \frac{1}{2}; \frac{z^2}{2}\Big) - \frac{\sqrt{2\pi} z + sin\left(\frac{\pi a}{z}\right)}{\Gamma\left(1-\frac{p}{2}\right)} \Omega\Big(\frac{1-p}{2}, \frac{3}{2}; \frac{z^2}{2}\Big) \right\}, \tag{2.14}$$

Ω is the confluent hypergeometric function (Abramowitz and Stegun, 1964) given by

$$\Omega(z;\delta,\epsilon) = 1 + \frac{\delta}{\epsilon}\frac{z}{1!} + \frac{\delta(\delta+1)}{\epsilon(\epsilon+1)}\frac{z^2}{2!} + \frac{\delta(\delta+1)(\delta+2)}{\epsilon(\epsilon+1)(\epsilon+2)}\frac{z^3}{3!} + \cdots, \tag{2.15}$$

and κ is a normalising constant given by

$$\kappa = \frac{2^\lambda \lambda}{\gamma\sqrt{\pi}} \Gamma\left(\lambda + \frac{1}{2}\right), \tag{2.16}$$

and λ and γ are parameters referred to as the shape and scale respectively.

If λ and γ are such that $\xi = \frac{\sqrt{2\lambda}}{\gamma}$, the NEG distribution would converge to the DE distribution having ξ parameter (Hoggart et al., 2008). Figure 2.1 shows the log density of the DE distribution and different three log densities of the NEG having the same density at zero. It can be seen that as λ decreases the shrinkage of large non zero coefficients in the NEG distribution is less than the DE distribution, because its tails are heavier than the DE density tails.

2.3.2 The optimisation algorithm

Hoggart et al. (2008) maximised the posterior density $f(\boldsymbol{\beta} \mid \boldsymbol{X}, \boldsymbol{y})$ over $\boldsymbol{\beta}$, where $\boldsymbol{X}$ is the $n \times p$ matrix containing the normalised genotype data, in which x_{ij} represents the jth genotype in person i and $\boldsymbol{y}$ is the vector of response variables which are coded 1 for a case and 0 for a

control. Using Bayes theorem the log posterior density is given by

$$\log f\left(\boldsymbol{\beta} \mid \boldsymbol{X}, \boldsymbol{y}\right) = \ell\left(\boldsymbol{y} \mid \boldsymbol{X}, \boldsymbol{\beta}\right) - \log f\left(\boldsymbol{\beta}\right) + \text{constant}, \tag{2.17}$$

where ℓ refers to the log-likelihood for the logistic regression model and $\log f\left(\boldsymbol{\beta}\right)$ refers to the negative log-prior density. Note that the minus sign here represents $\log\, f\left(\boldsymbol{\beta}\right)$ as a log penalty function. This is equivalent to maximising a penalised log-likelihood. If the DE distribution is selected as a prior, the log posterior density $\log f\left(\boldsymbol{\beta} \mid \boldsymbol{X}, \boldsymbol{y}\right)$ can be maximised over $\boldsymbol{\beta}$ using the Lasso procedure. Rather than using the EM algorithm which does not converge quickly with case-control data, Hoggart et al. (2008) applied the CIG algorithm (Bazaraa et al., 2013), which can optimise each coefficient in turn until convergence. No one had applied this algorithm with the NEG prior previously.

Since we want to find β_i that maximises the posterior distribution (i.e. is the posterior mode) we want to solve

$$\ell'\left(\beta_j\right) - \log\, f'\left(\beta_j\right) = 0. \tag{2.18}$$

Therefore Newton-Raphson's method can be used in this case by using the following expression

$$\beta_j^{(i+1)} = \beta_j^{(i)} - \frac{\ell'\left(\beta_j\right) - \log\, f'\left(\beta_j\right)}{\ell''\left(\beta_j\right) - \log\, f''\left(\beta_j\right)}, \tag{2.19}$$

where each derivative is with respect to β_j. $\log\, f'\left(\beta_j\right)$ is the first derivative of the log penalty function given by

$$\log\, f'\left(\beta_j\right) = \frac{\text{sign}\left(\beta_j\right)\left(2\lambda + 1\right)}{\gamma} \frac{D_{-(2\lambda+2)}\left(\frac{|\beta_j|}{\gamma}\right)}{D_{-(2\lambda+1)}\left(\frac{|\beta_j|}{\gamma}\right)}, \tag{2.20}$$

$\log f''(\beta_j)$ the second derivative of the log penalty function given by

$$\log f''(\beta_j) = \frac{4}{\gamma^2}\left((\lambda+1)\left(\lambda+\frac{1}{2}\right)\frac{D_{-(2\lambda+3)}\left(\frac{|\beta_j|}{\gamma}\right)}{D_{-(2\lambda+1)}\left(\frac{|\beta_j|}{\gamma}\right)} - \left(\left(\lambda+\frac{1}{2}\right)\frac{D_{-(2\lambda+2)}\left(\frac{|\beta_j|}{\gamma}\right)}{D_{-(2\lambda+1)}\left(\frac{|\beta_j|}{\gamma}\right)}\right)^2\right) \quad (2.21)$$

$\ell'(\beta_j)$ is the first derivative of the log-likelihood function given by

$$\ell'(\beta_j) = \frac{\partial}{\partial\beta_j}\ell(\beta_j) = \frac{\partial}{\partial\beta_j} - \sum_{i=1}^{n}\log\{1+\exp(\eta_i)\} = \sum_{i=1}^{n}\frac{x_{ij}y_i}{1+\exp(\eta_i)}, \quad (2.22)$$

and $\ell''(\beta_j)$ is the second derivative of the log-likelihood function given by

$$\ell''(\beta_j) = \frac{\partial^2}{\partial\beta_j^2}\ell(\beta_j) = \frac{\partial^2}{\partial\beta_j^2} - \sum_{i=1}^{n}\log\{1+\exp(\eta_i)\} = -\sum_{i=1}^{n}x_{ij}^2\frac{\exp(\eta_i)}{1+\exp(\eta_i)}, \quad (2.23)$$

where $\eta_j = \beta_0 + \sum_{j=1}^{p}\beta_{ij}x_{ij}$ and y_i refers to case-control status. Note that the log likelihood for the logistic regression model is given by

$$\ell(\boldsymbol{\beta}; \boldsymbol{X}, \boldsymbol{y}) = \log f(\boldsymbol{y} \mid \boldsymbol{X}, \boldsymbol{\beta}) = -\sum_{i=1}^{n}\log\{1+\exp(-\eta_i)\}. \quad (2.24)$$

If $\beta_j = 0$ then to update β_j one can calculate the limits of β_j as it approaches to zero from above and below. Thus the updating is not rejected if the β_j^{i+1} is not zero. β_j is non-zero if the following holds (Hoggart et al., 2008)

$$\mid \ell'(\beta_j = 0) \mid > \log f'\left(\beta_j = 0^+\right). \quad (2.25)$$

Calculating the first derivative of log posterior ℓ' could be computationally expensive, therefore the upper and lower bounds work as critical points in calculating ℓ'. If the absolute values of the upper bound or lower bound is greater than the penalty then calculating ℓ' is required.

The upper and lower bounds can be given as

$$-\frac{\sum_{i=1}^{n} I\left(y_i x_{ij} < 0\right) \mid x_{ij} \mid}{1+\exp\left(\eta_{\min}\right)} + \frac{\sum_{i=1}^{n} I\left(y_i x_{ij} > 0\right) \mid x_{ij} \mid}{1+\exp\left(\eta_{\max}\right)} < \frac{\partial}{\partial \beta_j} \log f\left(\boldsymbol{\beta} \mid \boldsymbol{X}, \boldsymbol{y}\right) \quad (2.26)$$
$$< -\frac{\sum_{i=1}^{n} I\left(y_i x_{ij} < 0\right) \mid x_{ij} \mid}{1+\exp\left(\eta_{\max}\right)} + \frac{\sum_{i=1}^{n} I\left(y_i x_{ij} > 0\right) \mid x_{ij} \mid}{1+\exp\left(\eta_{\min}\right)},$$

where $I\left(E\right)$ is the indicator function, which equals one if E holds and is zero otherwise.

2.3.3 Selecting prior parameters by controlling type I error

The shape and scale parameters define the penalty and if the data are standardised, the type I error rate (α) can be appropriately controlled by choosing the prior parameters. The key relationship between α and the prior is

$$\log f'(\beta = 0^+) = \sqrt{\frac{n_0 n_1}{n_0 + n_1}} \Phi^{-1}(1 - \alpha/2), \quad (2.27)$$

where n_0 refers to the number of cases, n_1 represents the number of controls, Φ^{-1} refers to the inverse normal cumulative distribution function and α represents the family wise error. Derivation of equation (2.27) is given in the supplementary material of Hoggart et al. (2008). The user needs to specify two quantities to run the Hyper lasso: shape (λ) and penalty (the value of $\log f'(\beta = 0^+)$) in the Equation (2.27) or shape (λ) and scale (γ). If shape and penalty or shape and scale are given then the Hyper lasso can be run directly by specifying these parameters and the data in the software (Hoggart et al., 2008).

In order to calculate the penalty given type I error $= 0.05$, one can apply Equation (2.27). Note that $f'(\beta = 0^+)$ represents the penalty of the prior; α in Equation (2.27) is the family wise error obtained after some sort of Bonferroni type adjustment for number of the parameters. Additionally, to calculate the scale of the prior for the given values, one can equate equations (2.20) and (2.27) and solve them in terms of the scale γ. In case of $\boldsymbol{\beta} = 0$ the scale

γ can be given as

$$\begin{aligned} \gamma &= \frac{\text{sign}(\boldsymbol{\beta})(2\lambda+1)}{\log f'(\boldsymbol{\beta})} \frac{D_{-(2\lambda+2)}\left(\frac{|\boldsymbol{\beta}|}{\gamma}\right)}{D_{-(2\lambda+1)}\left(\frac{|\boldsymbol{\beta}|}{\gamma}\right)} \\ &= \frac{(2\lambda+1)}{\log f'(0)} \frac{D_{-(2\lambda+2)}(0)}{D_{-(2\lambda+1)}(0)} \end{aligned} \tag{2.28}$$

The user needs to choose the inputs carefully, because some input configurations will cause the software to crash and this is one of the drawbacks of the HL software. HL only returns the posterior mode (log odds ratios) for those SNPs that are not shrunk towards zero.

2.3.4 Advantages and disadvantages

In this Section, we will focus on the advantages and disadvantages of the Hyper lasso method.

The normal-exponential-gamma prior lead to improved SNP selection compared to single-SNP tests (Hoggart et al., 2008). Moreover, an explicit expression for the type I error of the method was obtained, allowing calibration without using permutation techniques. The method can be applied to quantitative phenotypes as well as case-control phenotypes. Plus, the method can be extended to look for interactions. Furthermore, the method can also be used for fine-mapping studies with dense SNP sets. It can also be used for additive, recessive and dominant effects. Additionally, the false positive rate is reduced compared to single-SNP studies because of reduced residual variation (known functional SNPs are included in the model when looking for new SNPs), and power and localisation are also improved (Hoggart et al., 2008) .

However, the major drawback is that the use of the posterior mode means there is a hard selection rule: SNPs are either included or not. It would be preferable to give a probability of inclusion. The hard decision rule means that small changes in genotype could lead to potentially large changes in the SNPs selected. Also important, from the user perspective, the program crashes for some combinations of input parameters.

2.4 PiMASS

PiMASS (Guan and Stephens, 2011) can be used in multi-SNP association analyses (GWAS or fine mapping studies) to measure the proportion of the variance of the phenotype explained by the genotype, to calculate the posterior distribution of the number of causal SNPs, and to calculate the marginal posterior inclusion probabilities of each SNP, a measure of the strength of the marginal association.

Guan and Stephens (2011) examined the potential of applying Bayesian Variable Selection Regression (BVSR) to GWAS involving hundreds of thousands of covariates and thousands of observations, with a view to improving over previous GWAS analyses in the literature which relied on single-SNP analyses or penalised regression approaches such as the LASSO. Markov Chain Monte Carlo (MCMC) is used to implement the BVSR. The BVSR method is compared and contrasted with both the LASSO method and single-SNP methods.

In this Section, we will explain the method used in the PiMASS software, describe the prior and the method of updating the parameters.

2.4.1 Model and priors used

Although our data are case-control data, we will discuss the prior in the context of the multiple linear regression model, because the standard normal regression model was demonstrated in Guan and Stephens (2011). The software of PiMASS also allows the logistic likelihood that is appropriate for case-control data.

Guan and Stephens (2011) applied a standard normal linear regression model as follows

$$\boldsymbol{y}|\boldsymbol{\mu}, \boldsymbol{\beta}, \boldsymbol{X}, \tau \sim N_n\left(\boldsymbol{\mu} + \boldsymbol{X}\boldsymbol{\beta}, \tau^{-1}\boldsymbol{I}_n\right), \tag{2.29}$$

where $\boldsymbol{y}$ represents the vector response variable and $\boldsymbol{X}$ is the design matrix containing n individuals and p SNPs and $\boldsymbol{\beta}$ refers to the vector of regression coefficients. $\boldsymbol{\mu}$ represents the mean vector, τ refers to the precision of the errors. $\boldsymbol{I}_n$ represents the identity matrix with n rows and n columns and $N_n(.)$ refers to the multivariate normal distribution with n dimensions. Since most of the β_i will be zero, Guan and Stephens (2011) created $\boldsymbol{\gamma}$ as a p-

vector of indicator variables that indicates which elements of $\boldsymbol{\beta}$ are non zero. They let $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ be the parameter vector for those elements of $\boldsymbol{\gamma}$ that are 1 (i.e are included in the model) and $\boldsymbol{X}_{\boldsymbol{\gamma}}$ be the columns of $\boldsymbol{X}$ corresponding to those elements of $\boldsymbol{\gamma}$ that are 1.

$$\boldsymbol{y}|\boldsymbol{\gamma}, \boldsymbol{\mu}, \tau, \boldsymbol{\beta}, \boldsymbol{X}, \sim N_n(\boldsymbol{\mu} + \boldsymbol{X}_{\boldsymbol{\gamma}}\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \tau^{-1}\boldsymbol{I}_n). \tag{2.30}$$

The following priors were specified

$$\tau \sim Ga\left(\lambda/2, \kappa/2\right), \tag{2.31}$$

$$\mu|\tau \sim N\left(0, \sigma_\mu^2/\tau\right), \tag{2.32}$$

$$\gamma_j \sim Bernoulli\left(\pi\right), \tag{2.33}$$

$$\boldsymbol{\beta}_{\boldsymbol{\gamma}}|\tau, \boldsymbol{\gamma} \sim N_{|\boldsymbol{\gamma}|}\left(0, \left(\sigma_a^2/\tau\right)\boldsymbol{I}_{|\boldsymbol{\gamma}|}\right), \tag{2.34}$$

$$\boldsymbol{\beta}_{-\boldsymbol{\gamma}}|\boldsymbol{\gamma} \sim \delta_0, \tag{2.35}$$

where $|\boldsymbol{\gamma}| = \sum_j \gamma_j$ and $\boldsymbol{\beta}_{-\boldsymbol{\gamma}}$ represents the vector of coefficients where $\gamma_j = 0$ and δ_0 refers to a point mass on 0. The hyper-parameters are $\pi, \sigma_a, \lambda, \kappa$ and σ_μ .

The hyper-parameters of π and σ_a are important because π refers to the sparsity of the model (prior inclusion probability) and σ_a represents the typical sizes of the log odds ratios. The hyper-parameters v, κ and σ_μ^2 were considered less important than π and σ_a.

A logarithmic prior was used on the sparsity hyper-parameter π as follows

$$\log\pi \sim U(a, b), \tag{2.36}$$

where $a = \log(1/p)$ and $b = \log(M/p)$, so that $\frac{1}{p} \leq \pi \leq \frac{M}{p}$, where M is the number of potential SNPs in the model and p the total numbers of SNPs under consideration. They specified the maximum of M to be 400, because it can be expensive computationally if one were to select more than 400. The rationale behind the choice of a prior on $\log\pi$ rather than π is that a uniform prior would put too much prior mass on prior proportions close to the upper limit.

The authors suggested to concentrate on the meaning of priors in teams of the proportion

of variance in $\boldsymbol{y}$ explained by $\boldsymbol{X}_{\boldsymbol{\gamma}}$ (PVE). In the previous selected priors for π and σ_a^2, it is assumed that π and σ_a^2 are independent random variables and this also means that $\boldsymbol{\gamma}$ and σ_a^2 are independent as well. This means that the PVE increases with more covariates. However, this assumption is relaxed in PiMASS so that the model could have a small number of covariates but a high PVE and vice versa. This is achieved by specifying a joint distribution for $v\left(\gamma, \sigma_a^2\right)$.

Guan and Stephens (2011) specified a prior for $\sigma_a^2 \mid \boldsymbol{\gamma}$ by selecting a flat prior for PVE which takes values between 0 and 1. In order to achieve this they let $V(\boldsymbol{\beta}, \tau)$ refer to the empirical variance of $\boldsymbol{X}\boldsymbol{\beta}$ relative to the residual variance $\left(\frac{1}{\tau}\right)$ as follows

$$V(\boldsymbol{\beta}, \tau) = \frac{\tau}{n} \sum_{i=1}^{n} \left(\boldsymbol{x}_i^T \beta\right)^2 = \frac{\tau}{n} \boldsymbol{\beta}^T \boldsymbol{X}^T \boldsymbol{X} \boldsymbol{\beta} \tag{2.37}$$

where $\boldsymbol{X}$ has been centred. Now the total proportion of variance in $\boldsymbol{y}$ explained by $\boldsymbol{X}$ for the regression coefficients $\boldsymbol{\beta}$ can be expressed as

$$PVE(\boldsymbol{\beta}, \tau) = \frac{\frac{V(\boldsymbol{\beta}, \tau)}{\tau}}{\frac{1}{\tau} + \frac{V(\boldsymbol{\beta}, \tau)}{\tau}} = \frac{V(\boldsymbol{\beta}, \tau)}{1 + V(\boldsymbol{\beta}, \tau)}. \tag{2.38}$$

They selected a prior for $\boldsymbol{\beta} \mid \tau$ such that the induced prior on $PVE(\boldsymbol{\beta} \mid \tau)$ is distributed as a uniform distribution. Next, the expected value of $V(\boldsymbol{\beta}, \tau)$ that depends on σ_a^2 can be used to obtain $v\left(\boldsymbol{\gamma}, \sigma_a^2\right)$ as follows

$$v\left(\boldsymbol{\gamma}, \sigma_a^2\right) = E\left[V(\boldsymbol{\beta}, \tau) \mid \boldsymbol{\gamma}, \sigma_a^2, \tau\right] = \sigma_a^2 \sum_{j:\gamma_j = 1} s_j, \tag{2.39}$$

where $s_j = \frac{1}{n} \sum_{i=1}^{n} x_{ij}^2$ represents the variance of SNP j. Then they defined

$$h\left(\boldsymbol{\gamma}, \sigma_a^2\right) = \frac{E\left[V(\boldsymbol{\beta}, \tau)\right]}{1 + E\left[V(\boldsymbol{\beta}, \tau)\right]} = \frac{v\left(\boldsymbol{\gamma}, \sigma_a^2\right)}{1 + v\left(\boldsymbol{\gamma}, \sigma_a^2\right)}, \tag{2.40}$$

where h represents an approximation to the expectation of PVE for a given value for $\boldsymbol{\gamma}$ and σ_a^2. Also, they suggested a uniform prior on h that is independent of $\boldsymbol{\gamma}$. This leads to a prior

for $\sigma_a^2 \mid \boldsymbol{\gamma}$ as follows

$$\sigma_a^2(h,\gamma) = \frac{h\left(\boldsymbol{\gamma},\sigma_a^2\right)}{1-h\left(\boldsymbol{\gamma},\sigma_a^2\right)}\frac{1}{\sum_{j:\gamma_j=1} s_j}, \tag{2.41}$$

which achieves the goal of setting a uniform prior on PVE. The important feature for the σ_a^2 prior in the PiMASS method is that it applies less shrinkage than other priors previously used for σ_a^2.

2.4.2 Overview of MCMC used in PiMASS

In this section, we will describe the computation of the posterior distributions for the parameters in PiMASS.

For the MCMC computations the model was parametrised in terms of $(h,\boldsymbol{\gamma})$ instead of $(\sigma_a,\boldsymbol{\gamma})$ and the samples were obtained from the posterior distribution of $(h,\pi,\boldsymbol{\gamma})$ on the product space $(0,1)\times(0,1)\times\{0,1\}^p$ as follows

$$f\left(h,\pi,\boldsymbol{\gamma}\mid\boldsymbol{y}\right)\propto f\left(\boldsymbol{y}\mid h,\boldsymbol{\gamma}\right)f\left(h\right)f\left(\boldsymbol{\gamma}\mid\pi\right)f\left(\pi\right). \tag{2.42}$$

In order to sample from the posterior distributions of $h,\pi,\boldsymbol{\gamma}$, the marginal likelihood $f\left(\boldsymbol{y}\mid h,\boldsymbol{\gamma}\right)$ was calculated, which is possible because the parameters $\boldsymbol{\beta}$ and τ can be integrated out analytically. They considered the limit for the hyper-parameters v,κ as they approach 0 and σ_μ^2 as it approaches ∞ as was discussed in Section 2.4.1.

In the MCMC updating, a Metropolis-Hastings algorithm (Metropolis et al., 1953; Hastings, 1970) was applied to update update $h,\pi,\boldsymbol{\gamma}$ jointly. Here, local proposal distributions for h,π and $\boldsymbol{\gamma}$ were used. For the proposal distribution for $\boldsymbol{\gamma}$, covariates were added to the model based on their association with the phenotype $\boldsymbol{y}$. They implemented three proposal distribution for $\boldsymbol{\gamma}$, π and h. Firstly, a proposal distribution for $\boldsymbol{\gamma}$ can be implemented allowing for three possibilities: adding a covariate, removing a covariate and switching a covariate in the model. The proposal distribution for π is Beta$\left(\mid\boldsymbol{\gamma}'\mid,p-\mid\boldsymbol{\gamma}'\mid+1\right)$, where $\boldsymbol{\gamma}'$ is the proposed value of $\boldsymbol{\gamma}$. The proposal distribution of h can be given by adding a $U\left(-0.1,0.1\right)$ random variable to the current value of h. Moreover, in order to improve the convergence rate

of the MCMC the technique called "small world proposal" is applied. This technique compounds some local moves at random to make a longer range proposal. Finally, they applied Rao-Blackwellization techniques to reduce the posterior variance (see below).

Posterior inclusion probabilities via Rao-Blackwellisation

Identifying the covariates that have a high probability of inclusion in the model is an important aspect of inference in GWAS or fine-mapping studies. Therefore, the posterior inclusion probability (PIP) of the jth covariate $\mathrm{P}(\gamma_j = 1|\boldsymbol{y})$ is required. It can be calculated by counting the proportion of MCMC samples for which $\gamma_j = 1$, but this estimator might have a high sampling variance. For example, if one runs MCMC several times and each time calculates the proportion of samples for which $\gamma = 1$, the proportion might differ greatly. Rao-Blackwellised estimates are used in order to improve precision as follows

$$\mathrm{P}(\gamma_j = 1|\boldsymbol{y}) \approx (1/M) \sum_{i=1}^{M} \mathrm{P}\left(\gamma_j = 1|\boldsymbol{y}, \boldsymbol{\gamma}_{-j}^{(i)}, \boldsymbol{\beta}_{-j}^{(i)}, \tau^{(i)}, h^{(i)}, \pi^{(i)}\right), \tag{2.43}$$

where $\boldsymbol{\gamma}^{(i)}, \boldsymbol{\beta}^{(i)}, \tau^{(i)}, h^{(i)}, \pi^{(i)}$ denote the ith MCMC sample from the posterior distribution of these parameters given $\boldsymbol{y}$, and $\boldsymbol{\gamma}_{-j}$ and $\boldsymbol{\beta}_{-j}$ denote the vectors $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$ excluding the jth SNP. The probabilities that are being averaged here essentially involve simple univariate regressions of residuals against covariate j.

2.4.3 Applying PiMASS to case control data

Previously we discussed PiMASS with quantitative phenotype following a normal distribution. However, PiMASS can also be applied to a binary phenotype. To deal with this kind of data, the probit link function is commonly used. Latent variable $\boldsymbol{z}$ is used in practice by assuming that it follows the standard linear regression given in Equation (2.30) and relating $\boldsymbol{z}$ to $\boldsymbol{y}$ using this expression $y_i = 1(z_i > 0)$ (Albert and Chib, 1993). MCMC is performed to integrate out $\boldsymbol{z}$ for posterior inference. There is extra updating in the binary phenotype compared to the quantitative phenotype, that is updating the $\boldsymbol{z}$ variables. The advantage of the probit link function that is used for these latent Gaussian variables does not only allow us

to use the same priors for the quantitative phenotype, but also allows us to obtain the signal summaries by estimating PVE for the latent variables. Note that for a binary phenotype the priors still relate to unobserved latent Gaussian variables instead of the observed binary phenotype. Nevertheless, setting τ to be an improper prior for τ will result an improper posterior for both τ and $\boldsymbol{z}$. This issue could be solved by fixing τ to be 1 as has been done in (Albert and Chib, 1993). Alternatively, the latent variables $\boldsymbol{z}$ can be restricted to have variance 1. In order to improve the mixing, it can approximate the marginal distribution for $\boldsymbol{z}$ to be normal distribution to compatible with the standard linear regression given in Equation (2.30).

2.4.4 Advantages and disadvantages

Guan and Stephens (2011) successfully demonstrated BVSR as a competitive alternative to LASSO for GWAS, outperforming LASSO in predictive performance and also providing posterior probabilities for each relevant covariate. Also, the use of BVSR allowed the extraction of more information from the signal data as compared to single-SNP analysis.

However, PiMASS does not attempt to account for non-additive combinations of SNPs or SNPs that only have an effect in combination with others. Moreover, the method is susceptible to false positive associations when data quality is poor and genotype errors are correlated with phenotype, for example in a case-control study if the DNA quality differs substantially between the controls and cases. Unlike single-SNP studies, this error can impact association results at other SNPs. Also, the variables are assumed to be included in the model independently with equal probability π which ignores local spatial dependence of γ. This does not account for the common possibility of multiple functional variants appearing in a single gene (Guan and Stephens, 2011).

2.5 Normal-Gamma prior

In this Section, the normal gamma prior is discussed briefly and we leave detailed discussion of this prior to the next Chapter.

The method aims to select variables by shrinking "non significant" variables toward zero. This prior is one of the scale mixture of normals and it can be written in a hierarchical structure

as follows

$$\begin{aligned} \beta_i|\Psi_i &\sim N\left(0,\Psi_i\right) && (2.44)\\ \Psi_i &\sim \text{Ga}\left(.,.\right), && (2.45) \end{aligned}$$

where Ψ_i is the variance of the i regression coefficient. Because its tail is heavier than the normal distribution it would allow greater discrepancy in the sizes of regression coefficients comparing to a normal prior.

Griffin and Brown (2010) suggested this prior and they selected a particular prior for each parameter to construct the hierarchical form of the NG prior

$$\begin{aligned} \beta_i|\psi_i &\sim N\left(0,\psi_i\right), && (2.46)\\ \psi_i|\lambda,\gamma^{-2} &\sim Ga\left(\lambda,\frac{1}{2\gamma^2}\right), && (2.47)\\ \gamma^{-2}|\lambda &\sim Ga\left(2,\frac{M}{2\lambda}\right), && (2.48)\\ \lambda &\sim Ex(1), && (2.49) \end{aligned}$$

where $M=\frac{1}{p}\sum_{i=1}^{p}\hat{\beta}_i^2$, and $\hat{\beta}$ is the least square estimate for β.

2.6 Discussion

In this Thesis we choose the multivariate logistic regression and the Hyper lasso (HL), the PiMASS and the normal gamma (NG) prior, because of the properties that these methods have. The multivariate logistic regression was considered because it is the most common method used in fine-mapping studies. However, the HL prior was considered because it is the most common method that uses a Bayesian type approach and it can work with highly correlated data as in our case. In addition, PiMASS was selected because of its ability to deal with correlated data and because it provides a posterior inclusion probability. It appears to have been rarely used in reported association studies. Finally, the NG prior was selected because of its different prior structure that has never been used in fine-mapping studies before. The approach we took with the NG was not to update the SNPs in the model at each MCMC

iteration (like PiMASS) but just to update all the posterior effect sizes. This has the advantage of not having to worry about which SNPs to include but has the disadvantage of working with a covariance matrix with highly correlated explanatory variables. We wanted to assess how well it would work in this challenging setting. All the above methods perform variable selection in different ways.

Table 2.1 shows a summary of all four methods in terms of response, model, estimation method, estimator and prior. It can be seen that all methods have a binary response and logistic model except the NG prior which has continuous response and linear model because we use an asymptotic normal likelihood approximation instead of the logistic likelihood. SLR does not have a prior because it is a frequentist approach. However, the priors of the HL, PiMASS and the NG prior are Normal Exponential Gamma (NEG), a mixture of point mass at zero and Normal slab and Normal Gamma (NG) respectively. In order to compare between the four methods, we calculate the smallest initial univariate p-value threshold that select potential causal SNPs for SLR using stepwise logistic regression. We used different posterior summaries in the Bayesian approaches: HL, PiMASS and the NG prior to calculate posterior modes, posteriori inclusion probabilities and credible intervals respectively.

Method	SLR	Hyper lasso	PiMASS	NG
Response	Binary	Binary	Binary	Continuous
Model	Logistic	Logistic	Logistic	Linear
Estimation method	Stepwise logistic regression	Shrinkage regression	Shrinkage regression	Shrinkage regression
Estimator	Initial p-value	Posterior Modes	Posteriori inclusion probabilities	Credible interval size
Prior	None	NEG	Normal	NG

Table 2.1: Summary of the four methods applied in this research.

Chapter 3

Normal gamma prior

In the previous Chapter we discussed some multivariate statistical model selection approaches both frequentist (Multivariate Logistic Regression, Step-wise Logistic Regression) and Bayesian (HL, PiMASS and NG prior) for case-control data that are used in fine-mapping studies.

In this Chapter, we will study the normal gamma (NG) prior and investigate its features in the Bayesian framework. Additionally, an asymptotic Gaussian likelihood will be used here rather than the true logistic likelihood, and the full conditional distribution for each parameter will be calculated. Some specific computational tools will be used in the MCMC updating when estimating the parameters and their use will be described. Lastly, we will also use some prior information about likely effect sizes to inform our prior parameters. In order to find out the "best" rate parameter κ for the hyperparameter of the NG prior (λ), we will use the published "top hits" from large breast cancer genome wide association studies (Michailidou et al., 2015; Fachal and Dunning, 2015).

Here we use the NG prior to mean a fully Bayesian inference procedure with a particular prior (NG prior) with Gaussian likelihood. Griffin and Brown (2010) only states the full conditional distributions However, we will derive the full conditional distributions for the new likelihood and correct some errors in their calculations. We also need to modify the calculations for our asymptotic likelihood.

In Bayesian regression analysis, choosing the appropriate prior for the model parameters plays an important role. There are many suitable different distributions which can be selected as the prior for the unknown regression parameters, such as the double exponential distribu-

tion in Hyper lasso (Hoggart et al., 2008), the spike and slab in PiMASS (Guan and Stephens, 2011) and the normal gamma distribution (Griffin and Brown, 2010). There is existing software for both Hyper lasso and PiMASS but for implementing the NG prior there is only code in Matlab. However, we wrote R code from scratch to deal with the different likelihood. These three methods have different procedures. Hyper lasso shrinks many SNPs to exactly zero (posterior mode), but the NG prior shrinks SNPs towards zero but not exactly to zero. In PiMASS during each iterations there is a hard decision to select SNPs into the model and these SNPs only are potentially shrunk towards zero. Therefore, PiMASS is somewhere in between Hyper lasso and the NG prior.

Griffin and Brown (2010) investigated the influence of the NG prior on β on the estimated posterior distribution in terms of the posterior mean and variance. They demonstrated the relationship between the posterior mean, which is a shrinkage estimator, and how the level of the shrinkage can be affected by the prior distribution used. Additionally, they generalised the double exponential prior distribution commonly used in Bayesian regression problems.

The standard multiple linear regression model is

$$\boldsymbol{y} = \alpha \mathbf{1}_n + \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{3.1}$$

where $\boldsymbol{y}$ refers to the a vector of response variables, $\boldsymbol{\epsilon}$ is a vector of error terms and ϵ_is are independent and distributed as normal with mean 0 and variance σ^2, and $\boldsymbol{X}$ is a matrix of independent variables with n rows and p columns. α refers to an intercept and $\mathbf{1}_n$ is a vector of 1s with n elements. $\boldsymbol{\beta}$ is a p-vector of the regression of coefficients in the model. Analysing properties of the NG prior within the context of the above model within the Bayesian framework was the aim of Griffin and Brown (2010). They were concerned with selecting a suitably flexible prior for the regression coefficient β that considerably shrunk small parameter estimates but applied less shrinkage to large parameter estimates.

The scale mixture of normals distribution (see for example Andrews and Mallows (1974); West (1987)) is a common prior distributions for the regression coefficients. This distribution

can be expressed as follows

$$\pi\left(\beta_i\right) = \int N\left(\beta_i|0,\psi_i\right) dF\left(\psi_i\right), \tag{3.2}$$

where F represents a mixing distribution. Griffin and Brown (2010) proposed a continuous prior distribution which has the following hierarchical structure

$$\beta_i|\psi_i \sim N\left(0,\psi_i\right) \text{ and } \psi_i \sim Ga\left(\theta,\phi\right), \tag{3.3}$$

where ψ_i is the variance of the ith regression coefficients, so that each β_i has a potentially different variance ψ_i a priori. They set the prior for $\psi_i|\lambda,\gamma^2$ to be distributed as follows

$$\psi_i|\lambda,\gamma^2 \sim Ga\left(\lambda,\frac{1}{2\gamma^2}\right), \tag{3.4}$$

which allows sufficient flexibility in the variances. This seems appropriate for fine-mapping where most coefficients are zero or close to zero but we want to allow for the possibility of large effect sizes. From Equations 3.3 and 3.4, the prior for $\beta_i|\lambda,\gamma^2$ takes a closed form via the following integral

$$\begin{aligned}
\pi\left(\beta_i|\lambda,\gamma^2\right) &= \int_0^\infty \pi\left(\beta_i|\psi_i\right)\ \pi\left(\psi_i|\lambda,\gamma^2\right)\ d\psi_i \\
&= \int_0^\infty \frac{1}{\sqrt{2\pi\psi_i}}\, e^{-\frac{\beta_i^2}{2\psi_i}}\, \frac{\left(\frac{1}{2\gamma^2}\right)^\lambda}{\Gamma\left(\lambda\right)}\, \psi_i^{\lambda-1}\, e^{-\frac{\psi_i}{2\gamma^2}}\, d\psi_i \\
&= \frac{1}{\sqrt{2\pi}}\, \frac{\left(\frac{1}{2\gamma^2}\right)^\lambda}{\Gamma\left(\lambda\right)} \int_0^\infty \psi_i^{-\frac{1}{2}}\, e^{-\frac{\beta_i^2}{2\psi_i}}\, \psi_i^{\lambda-1}\, e^{-\frac{\psi_i}{2\gamma^2}}\, d\psi_i \\
&= \frac{\left(\frac{1}{2\gamma^2}\right)^\lambda}{\sqrt{2\pi}\,\Gamma\left(\lambda\right)} \int_0^\infty \psi_i^{\left(\lambda-\frac{1}{2}\right)-1}\, e^{-\frac{1}{2}\left(\frac{1}{2\gamma^2}\,\psi_i+\frac{\beta_i^2}{\psi_i}\right)}\, d\psi_i \\
&= \frac{\left(\frac{1}{2\gamma^2}\right)^\lambda}{\sqrt{2\pi}\,\Gamma\left(\lambda\right)}\, \frac{2\, K_{\left(\lambda-\frac{1}{2}\right)}\left(\sqrt{\frac{\beta_i^2}{\gamma^2}}\right)}{\left(\frac{1}{\gamma^2\,\beta_i^2}\right)^{\frac{\lambda-\frac{1}{2}}{2}}} \int_0^\infty \frac{\left(\frac{1}{\gamma^2\,\beta_i^2}\right)^{\frac{\lambda-\frac{1}{2}}{2}}}{2\, K_{\left(\lambda-\frac{1}{2}\right)}\left(\sqrt{\frac{\beta_i^2}{\gamma^2}}\right)}\, \psi_i^{\left(\lambda-\frac{1}{2}\right)-1}\, e^{-\frac{1}{2}\left(\frac{1}{2\gamma^2}\,\psi_i+\frac{\beta_i^2}{\psi_i}\right)}\, d\psi_i
\end{aligned}$$

$$
\begin{aligned}
&= \frac{\left(\frac{1}{2\gamma^2}\right)^{\lambda}}{\sqrt{2\pi}\,\Gamma(\lambda)} \frac{2\,K_{\left(\lambda-\frac{1}{2}\right)}\left(\sqrt{\frac{\beta_i^2}{\gamma^2}}\right)}{\left(\frac{1}{\gamma^2\,\beta_i^2}\right)^{\frac{\lambda-\frac{1}{2}}{2}}} \\
&= \frac{1}{\sqrt{\pi}\;2^{\lambda-\frac{1}{2}}\;\gamma^{\lambda+\frac{1}{2}}\Gamma\lambda} \,|\beta_i^{\lambda-\frac{1}{2}}\mid\; K_{\left(\lambda-\frac{1}{2}\right)}\left(\frac{|\,\beta_i|}{\gamma}\right), \qquad (3.5)
\end{aligned}
$$

where K refers to the modified Bessel function of the third kind (Abramowitz and Stegun, 1964) and

$$
\frac{\left(\frac{1}{\gamma^2\,\beta_i^2}\right)^{\frac{\lambda-\frac{1}{2}}{2}}}{2\,K_{\left(\lambda-\frac{1}{2}\right)}\left(\sqrt{\frac{\beta_i^2}{\gamma^2}}\right)}\;\psi_i^{\left(\lambda-\frac{1}{2}\right)-1}\;e^{-\frac{1}{2}\left(\frac{1}{2\gamma^2}\;\psi_i+\frac{\beta_i^2}{\psi_i}\right)} \qquad (3.6)
$$

is the generalised inverse Gaussian (GIG) probability density function, with support $[0,\infty)$. The variance of $\beta_i \mid \lambda,\gamma^2$ is calculated using the law of the total variance see for example Casella and Berger (2002). It is calculated as follows

$$
\begin{aligned}
\boldsymbol{Var}\left(\beta_i \mid \lambda,\gamma^2\right) &= \mathbb{E}_{\psi_i}\left(\boldsymbol{Var}_{\beta_i|\psi_i}\left(\beta_i \mid \psi_i\right)\right) + \boldsymbol{Var}_{\psi_i}\left(\mathbb{E}_{\beta_i|\psi_i}\left(\beta_i \mid \psi_i\right)\right) \qquad (3.7)\\
&= \mathbb{E}\left(\psi_i \mid \lambda,\gamma^2\right) \qquad (3.8)\\
&= 2\lambda\gamma^2 \qquad (3.9)
\end{aligned}
$$

since ψ_i has a $Ga\left(\lambda,\frac{1}{2\gamma^2}\right)$ distribution.

Changing the shape of λ can clearly affect the marginal distribution of β_i. Figure 3.1 shows the marginal distribution for β_i for the normal gamma prior distribution and how it depends on the shape parameter of λ. It can be seen that as the shape λ decreases the mass located close to zero increases.

The value of parameters λ and γ^2 play an important role in the posterior parameter estimation. In the normal gamma prior a fully Bayesian method is applied by setting up a prior for each of these two parameters. These hyperparameters are difficult to estimate via an empirical Bayes method, and this is discussed by Park and Casella (2008). Griffin and Brown

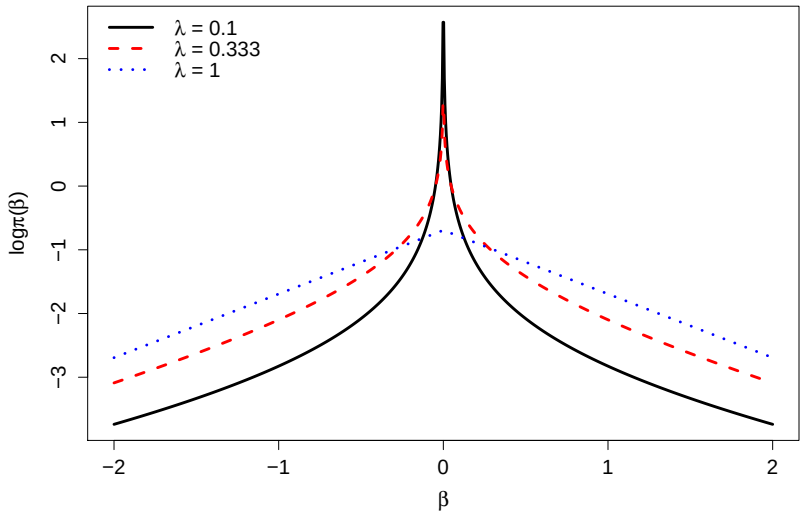


Figure 3.1: The log density of $\pi\left(\beta|\lambda,\gamma^2\right)$ *with var* $\left(\beta|\lambda,\gamma^2\right) = 2$ *for* $\lambda = 0.1$ *(solid line),* $\lambda = 0.333$ *(dashed line) and* $\lambda = 1$ *(dotted line). This plot is reproduced using the same values assumed in Griffin and Brown (2010).*

(2010) admit that their posterior distributions are assumed to be highly multi-modal. As a consequence of this we use the implementation in Equation (3.49) to explore the modes of the multi-modal posterior effectively.

Griffin and Brown (2010) utilised a particular prior for each of the model parameters. They selected the exponential distribution with mean 1 for λ, because it gives variability around the Bayesian Lasso prior with $\lambda = 1$. Additionally, they assumed the variance of β_i $(i.e.\ 2\lambda\gamma^2)$ is distributed as $IG\left(2, M\right)$. They also used a non-informative prior for the intercept parameter α. However, an improper prior can give an improper posterior; therefore, we propose using a normal prior with mean 0 and variance 0.1 as prior for α, because the intercept is likely to take a value close to zero.

Griffin and Brown (2010) selected a particular prior for each parameter to construct the hierarchical form of the NG prior. With the modification to α, it is given as follows

$$\alpha \sim N\left(0, 0.1\right), \tag{3.10}$$

$$\beta_i \mid \psi_i \sim N\left(0, \psi_i\right), \tag{3.11}$$

$$\psi_i \mid \lambda, \gamma^{-2} \sim Ga\left(\lambda, \frac{1}{2\gamma^2}\right), \tag{3.12}$$

$$\gamma^{-2} \mid \lambda \sim Ga\left(2, \frac{M}{2\lambda}\right), \tag{3.13}$$

$$\lambda \sim \text{Ex}(1), \tag{3.14}$$

where the prior of $\gamma^{-2}|\lambda$ is a transformation from the inverse gamma distribution of the prior of $2\lambda\gamma^2$ ($2\lambda\gamma^2 \sim IG(2, M)$) The distribution of $(2\lambda\gamma^2)^{-1}$ is the gamma distribution with scale 2 and scale M $\left((2\lambda\gamma^2)^{-1} \sim Ga(2, M)\right)$. Then, a further transformation yields

$$\gamma^{-2} \mid \lambda \sim Ga\left(2, \frac{M}{2\lambda}\right), \tag{3.15}$$

where $M = \frac{1}{p}\sum_{i=1}^{p} \hat{\beta}_i^2$, and $\hat{\beta}$ is the least squares estimate for β. Note that the expected value of $2\lambda\gamma^2$ is M, which is sensible since $2\lambda\gamma^2$ is the prior variance of β_i. Because of high levels of multi collinearity we choose to use $M = 1$ rather than a value based on the unstable estimates $\hat{\beta}_i$.

They applied both Gibbs sampling and Metropolis-Hasting updates to sample from the posterior distribution for all the parameters. To update the parameters they used the full conditional distribution for each parameter, and these distributions will be discussed in Section 3.2.

3.1 Asymptotic normal likelihood distribution

In classical statistical analysis, the maximum likelihood estimator is used widely because it has several desirable features. One of these is that it has an asymptotic normal distribution, which can be defined as follows: $\hat{\theta}$ is considered asymptotically normal if it holds that

$$\sqrt{n}\left(\hat{\theta} - \theta_0\right) \longrightarrow N\left(0, \sigma_{\theta_0}^2\right), \tag{3.16}$$

where $\sigma^2_{\theta_0}$ refers to the asymptotic variance of the estimator $\hat{\theta}$, θ_0 represents the unknown true parameter value and n is the sample size . This can be rewritten as follows

$$\hat{\theta} \sim N\left(\theta_0, \frac{\sigma^2_{\theta_0}}{n}\right). \tag{3.17}$$

Asymptotically, the estimator $\hat{\theta}$ convergences to the unknown parameter at a rate of $\frac{1}{\sqrt{n}}$ (Miller, 1977).

In our case, the estimator $\hat{\boldsymbol{\beta}}$ conditional on $\boldsymbol{\beta}$ is distributed as a normal distribution

$$\hat{\boldsymbol{\beta}}|\boldsymbol{\beta} \sim \mathrm{N}\left(\boldsymbol{\beta}, \boldsymbol{V}\right), \tag{3.18}$$

where $\boldsymbol{V}$ is the estimated variance-covariance matrix from the fitted generalised linear model (GLM).

For the preceding reasons, rather than using the correct logistic likelihood for the response which is computationally demanding, we will use the asymptotic Gaussian distribution for the maximum likelihood estimate for the model coefficients (log odds ratios). This enables us to speed up our MCMC analysis by using the Gaussian linear model framework with Gibbs updates for $\boldsymbol{\beta}$. The sample sizes used in fine-mapping make this a reasonable asymptotic approximation.

In order to avoid problems with exact multi collinearity we remove SNPs that have identical genotypes to another SNP. This allows estimation of $\boldsymbol{V}$ in the GLM.

3.2 Calculating full conditional distributions

In this section, we will calculate the full conditional distributions for each model parameter. These are modified versions of those given in Griffin and Brown (2010) based on using the asymptotic Gaussian likelihood and correcting some errors identified in Griffin and Brown (2010). The full conditional distribution of a particular parameter comes from the joint distribution of all parameters and the asymptotic normal likelihood distribution.

3.2.1 Joint distribution

We will calculate the joint distributions for the parameters and the response given by Griffin and Brown (2010) for the asymptotic normal likelihood. The joint distributions are given by

$$\begin{aligned} f\left(\alpha, \boldsymbol{\beta}, \boldsymbol{X}, \boldsymbol{\psi}, \lambda, \gamma^{-2}, \boldsymbol{y}\right) &= f(\alpha) \prod_{i=1}^{p} f(\boldsymbol{\beta} \mid \boldsymbol{\psi}) \prod_{i=1}^{p} f\left(\boldsymbol{\psi} \mid \lambda, \gamma^{-2}\right) \times f\left(\lambda, \gamma^{-2}\right) \times \\ &\quad f\left(\boldsymbol{y}; \boldsymbol{X}, \boldsymbol{\beta}, \boldsymbol{\psi}, \lambda, \gamma^{-2}\right) \\ &= f(\alpha) \prod_{i=1}^{p} f(\boldsymbol{\beta} \mid \boldsymbol{\psi}) \prod_{i=1}^{p} f\left(\boldsymbol{\psi} \mid \lambda, \gamma^{-2}\right) \times f\left(\gamma^{-2} \mid \lambda\right) \times f(\lambda) \times \\ &\quad f\left(\boldsymbol{y}; \boldsymbol{X}, \boldsymbol{\beta}, \boldsymbol{\psi}, \lambda, \gamma^{-2}\right). \end{aligned} \tag{3.19}$$

It is clear that four prior densities and the likelihood are required to calculate the joint distribution. These prior densities are from Equations 3.10 to 3.14,

$$f(\alpha) = \frac{1}{\sqrt{(2\pi)\,0.1}} \exp\left\{-\frac{1}{2}\frac{(\alpha)^2}{0.1}\right\} \tag{3.20}$$

$$\begin{aligned} f(\boldsymbol{\beta} \mid \boldsymbol{\Psi}) &= \frac{1}{\sqrt{(2\pi)^n |\boldsymbol{\Psi}|}} \exp\left\{-\frac{1}{2}(\boldsymbol{\beta})^T \boldsymbol{\Psi}^{-1}(\boldsymbol{\beta})\right\} \\ &= \frac{1}{\sqrt{(2\pi)^n |\psi_i|}} \exp\left\{-\frac{1}{2}\sum_{i=1}^{n}\frac{(\beta_i)^2}{\psi_i}\right\} \end{aligned} \tag{3.21}$$

$$\begin{aligned} f\left(\boldsymbol{\Psi} \mid \lambda, \gamma^{-2}\right) &= \prod_{i=1}^{p} \frac{\left(\frac{1}{2\gamma^2}\right)^{\lambda}}{\Gamma(\lambda)} (\boldsymbol{\Psi})^{\lambda-1} \exp\left\{-\frac{\boldsymbol{\Psi}}{2\gamma^2}\right\} \\ &= \frac{\left(\frac{1}{2\gamma^2}\right)^{p\lambda}}{(\Gamma(\lambda))^p} \prod_{i=1}^{p} (\psi_i)^{\lambda-1} \exp\left\{-\frac{\psi_i}{2\gamma^2}\right\} \end{aligned} \tag{3.22}$$

$$f\left(\gamma^{-2} \mid \lambda\right) = \frac{(M/2\lambda)^2}{\Gamma(2)} \left(\gamma^{-2}\right)^{2-1} \exp\left\{-\frac{M}{2\lambda}\gamma^{-2}\right\} \tag{3.23}$$

$$f(\lambda) = \exp(-\lambda), \tag{3.24}$$

where $M = 1$.

3.2.2 Full conditional distributions for α and $\boldsymbol{\beta}$

Here we will calculate the full conditional distributions for α and $\boldsymbol{\beta}$. Initially, let $\boldsymbol{\phi}$ be a vector that contains α and $\boldsymbol{\beta}$ as follows

$$\boldsymbol{\phi} = (\alpha, \boldsymbol{\beta})^T, \tag{3.25}$$

and let $\boldsymbol{\Lambda}$ be a diagonal matrix, where its diagonal elements are the precisions of α and $\boldsymbol{\beta}$ as follows

$$\boldsymbol{\Lambda} = \text{diag}\left(\frac{1}{0.1}, \frac{1}{\psi_1}, \frac{1}{\psi_2}, \ldots, \frac{1}{\psi_p}\right). \tag{3.26}$$

The full conditional distribution for $\boldsymbol{\phi}$ comes from the prior of $f(\boldsymbol{\phi})$, which is given by

$$f(\boldsymbol{\phi}) \propto \exp\left(\frac{-1}{2}\boldsymbol{\phi}^T\boldsymbol{\Lambda}\boldsymbol{\phi}\right), \tag{3.27}$$

and the likelihood $f\left(\hat{\boldsymbol{\phi}} \mid \boldsymbol{\phi}\right)$, which is given by

$$f\left(\hat{\boldsymbol{\phi}} \mid \boldsymbol{\phi}\right) \propto \exp\left\{-\frac{1}{2}\left(\hat{\boldsymbol{\phi}} - \boldsymbol{\phi}\right)^T \boldsymbol{V}^{-1}\left(\hat{\boldsymbol{\phi}} - \boldsymbol{\phi}\right)\right\}. \tag{3.28}$$

Expanding the terms in the exponent in Equation (3.28) gives

$$-\frac{1}{2}\left(\hat{\boldsymbol{\phi}}^T\boldsymbol{V}^{-1}\hat{\boldsymbol{\phi}} - \hat{\boldsymbol{\phi}}^T\boldsymbol{V}^{-1}\boldsymbol{\phi} - \boldsymbol{\phi}^T\boldsymbol{V}^{-1}\hat{\boldsymbol{\phi}} + \boldsymbol{\phi}^T\boldsymbol{V}^{-1}\boldsymbol{\phi}\right). \tag{3.29}$$

Equation (3.29) can be written as

$$\boldsymbol{\phi}^T\boldsymbol{V}^{-1}\boldsymbol{\phi} - 2\hat{\boldsymbol{\phi}}^T\boldsymbol{V}^{-1}\boldsymbol{\phi}, \tag{3.30}$$

because $\hat{\boldsymbol{\phi}}^T\boldsymbol{V}^{-1}\hat{\boldsymbol{\phi}}$ is a constant because it is not function of $\boldsymbol{\phi}$ and $\hat{\boldsymbol{\phi}}^T\boldsymbol{V}^{-1}\boldsymbol{\phi} = \boldsymbol{\phi}^T\boldsymbol{V}^{-1}\hat{\boldsymbol{\phi}}$, because the variance-covariance matrix $\boldsymbol{V}^{-1}$ is a symmetric matrix, and $\hat{\boldsymbol{\phi}}^T\boldsymbol{V}^{-1}\hat{\boldsymbol{\phi}}$ is a scalar

quantity. Hence the likelihood can be written as

$$f\left(\hat{\boldsymbol{\phi}} \mid \boldsymbol{\phi}\right) \propto \exp\left(-\frac{1}{2}\left[\boldsymbol{\phi}^T \boldsymbol{V}^{-1}\boldsymbol{\phi} - 2\hat{\boldsymbol{\phi}}^T \boldsymbol{V}^{-1}\boldsymbol{\phi}\right]\right). \tag{3.31}$$

Therefore, the posterior can be given by

$$f(\boldsymbol{\phi} \mid \boldsymbol{\Lambda}, \lambda, \gamma^{-2}, \hat{\boldsymbol{\phi}}) \propto \exp\left(-\frac{1}{2}\left[\boldsymbol{\phi}^T \boldsymbol{V}^{-1}\boldsymbol{\phi} - 2\hat{\boldsymbol{\phi}}^T \boldsymbol{V}^{-1}\boldsymbol{\phi}\right]\right) \exp\left(-\frac{1}{2}\boldsymbol{\phi}^T \boldsymbol{\Lambda}\boldsymbol{\phi}\right) \tag{3.32}$$

$$f(\boldsymbol{\phi} \mid \boldsymbol{\Lambda}, \lambda, \gamma^{-2}, \hat{\boldsymbol{\phi}}) \propto \exp\left\{-\frac{1}{2}\left[\boldsymbol{\phi}^T \left(\boldsymbol{V}^{-1} + \boldsymbol{\Lambda}\right)\boldsymbol{\phi} - 2\hat{\boldsymbol{\phi}}^T \boldsymbol{V}^{-1}\boldsymbol{\phi}\right]\right\}. \tag{3.33}$$

We are interested in expressing Equation (3.33) as a Gaussian kernel as follows

$$f(\boldsymbol{\phi} \mid \boldsymbol{\Lambda}, \lambda, \gamma^{-2}, \hat{\boldsymbol{\phi}}) \propto \exp\left\{-\frac{1}{2}\left(\boldsymbol{\phi} - \boldsymbol{\mu}\right)^T \boldsymbol{\Sigma}\left(\boldsymbol{\phi} - \boldsymbol{\mu}\right)\right\} \tag{3.34}$$

$$\propto \exp\left\{-\frac{1}{2}\left(\boldsymbol{\phi}^T\boldsymbol{\Sigma}\boldsymbol{\phi} - 2\boldsymbol{\mu}^T\boldsymbol{\Sigma}\boldsymbol{\phi} + \boldsymbol{\mu}^T\boldsymbol{\Sigma}\boldsymbol{\mu}\right)\right\}, \tag{3.35}$$

where $\boldsymbol{\mu}$ represents the posterior mean and $\boldsymbol{\Sigma}$ refers to an inverse variance-covariance matrix. To this end, we want to complete the square in Equation (3.33) to obtain the normal distribution for the posterior distribution (the full conditional distribution) as given in Equation (3.35). By looking at the Equation (3.33) we can observe appropriate values for the inverse variance-covariance matrix and the posterior mean are

$$\boldsymbol{\Sigma} = \boldsymbol{V}^{-1} + \boldsymbol{\Lambda} \tag{3.36}$$

$$\boldsymbol{\mu} = \boldsymbol{\Sigma}^{-1}\boldsymbol{V}^{-1}\hat{\boldsymbol{\phi}} \tag{3.37}$$

$$= \left(\boldsymbol{V}^{-1} + \boldsymbol{\Lambda}\right)^{-1}\boldsymbol{V}^{-1}\hat{\boldsymbol{\phi}} \tag{3.38}$$

This is checked in the following calculations. Let us substitute Equations (3.36) and (3.37) into the Equation (3.35). Then each term can be rewritten as follows

$$\text{The first term: } \boldsymbol{\phi}^T\boldsymbol{\Sigma}\boldsymbol{\phi} = \boldsymbol{\phi}^T\left(\boldsymbol{V}^{-1} + \boldsymbol{\Lambda}\right)\boldsymbol{\phi} \tag{3.39}$$

$$\text{The second term: } \boldsymbol{\mu}^T\boldsymbol{\Sigma}\boldsymbol{\phi} = \left(\boldsymbol{\Sigma}^{-1}\boldsymbol{V}^{-1}\hat{\boldsymbol{\phi}}\right)^T\boldsymbol{\Sigma}\boldsymbol{\phi} \tag{3.40}$$

$$= \hat{\boldsymbol{\phi}}^T\left(\boldsymbol{V}^{-1}\right)^T\left(\boldsymbol{\Sigma}^{-1}\right)^T\boldsymbol{\Sigma}\boldsymbol{\phi} \tag{3.41}$$

$$= \hat{\boldsymbol{\phi}}^T \boldsymbol{V}^{-1} \boldsymbol{\Sigma}^{-1} \boldsymbol{\Sigma} \boldsymbol{\phi} \tag{3.42}$$

$$= \hat{\boldsymbol{\phi}}^T \boldsymbol{V}^{-1} \boldsymbol{\phi}, \tag{3.43}$$

the third term $\left(\boldsymbol{\mu}^T \boldsymbol{\Sigma} \boldsymbol{\mu}\right)$ does not depend on $\boldsymbol{\phi}$. Therefore, we will end up with

$$f(\boldsymbol{\phi} \mid \boldsymbol{\Lambda}, \lambda, \gamma^{-2}, \hat{\boldsymbol{\phi}}) \sim \boldsymbol{N}\left(\left(\boldsymbol{V}^{-1} + \boldsymbol{\Lambda}\right)^{-1} \boldsymbol{V}^{-1} \hat{\boldsymbol{\phi}}, \boldsymbol{V}^{-1} + \boldsymbol{\Lambda}\right). \tag{3.44}$$

3.2.3 Full conditional distributions for ψ_i

In this section, we will calculate the full conditional distribution for ψ_i. The likelihood has no effect on the full conditional distribution of ψ_i. In this situation, the full condition can be derived from the joint distribution as follows

$$f(\psi_i \mid \beta_i, \lambda, \gamma^{-2}, y_i, \boldsymbol{X}) \propto \frac{1}{\sqrt{\psi_i}} \exp\left(-\frac{\beta_i^2}{2\psi_i}\right) (\psi_i)^{\lambda-1} \exp\left(-\frac{\psi_i}{2\gamma^2}\right) \tag{3.45}$$

$$= (\psi_i)^{-\frac{1}{2}} \exp\left(-\frac{\beta_i^2}{2\psi_i}\right) (\psi_i)^{\lambda-1} \exp\left(-\frac{\psi_i}{2\gamma^2}\right) \tag{3.46}$$

$$= (\psi_i)^{\lambda-\frac{1}{2}-1} \exp\left\{-\frac{1}{2}\left(\gamma^{-2}\psi_i + \frac{\beta_i^2}{2\psi_i}\right)\right\} \tag{3.47}$$

$$= (\psi_i)^{(\lambda-\frac{1}{2})-1} \exp\left\{-\frac{1}{2}\left(\gamma^{-2}\psi_i + \frac{\beta_i^2}{2\psi_i}\right)\right\}. \tag{3.48}$$

This expression is a kernel of generalised inverse Gaussian (GIG) distribution with $m = \lambda - \frac{1}{2}, c = \gamma^{-2}$ and $d = \beta_i^2$ (see section 3.3.2).

3.2.4 Full conditional distributions for λ

The multimodal posterior meant that Griffin and Brown (2010) had to choose an approach that would allow different modes to be explored. We will calculate the full conditional distribution for λ jointly with γ^{-2} to tackle this problem. It can be seen from Equation (3.19) that the likelihood has no effect on the full conditional distribution of λ. As a result, the full

conditional distribution of λ will be extracted from the joint distribution as follows

$$\begin{aligned} f(\lambda \mid \beta_i, \psi_i, \gamma^{-2}, y_i, \boldsymbol{X}) \quad &\propto \quad \frac{\left(\frac{1}{2\gamma_i^2}\right)^{p\lambda}}{\{\Gamma(\lambda)\}^p} \times \left(\prod_{i=1}^{p} \psi_i\right)^{\lambda-1} \times \prod_{i=1}^{p} \exp\left(-\frac{\psi_i}{2\gamma^2}\right) \times \\ &\quad \left(\gamma^{-2}\right)^{2-1} \times \left(\frac{M}{2\lambda}\right)^2 \times \exp\left(-\frac{M}{2\lambda}\gamma^{-2}\right) \times \\ &\quad \pi(\lambda) \\ &= \quad \frac{\left(\frac{1}{2\gamma^2}\right)^{p\lambda}}{\{\Gamma(\lambda)\}^p} \times \left(\prod_{i=1}^{p} \psi_i\right)^{\lambda-1} \times \prod_{i=1}^{p} \exp\left(-\frac{\psi_i}{2\gamma^2}\right) \times \\ &\quad \left(\frac{M^2}{2}\right) \times \left(\frac{1}{2\lambda\gamma^2}\right) \times \left(\frac{1}{\lambda}\right) \times \exp\left(-\frac{M}{2\lambda}\gamma^{-2}\right) \times \\ &\quad \pi(\lambda). \end{aligned}$$

Here λ is updated using a Metropolis-Hasting approach. $\lambda' = \exp(\sigma_\lambda^2 z)\,\lambda$ is chosen to be the proposal distribution, where z has a standard normal distribution. Also, we choose an adjustment value for γ'^2 as follows

$$\gamma'^2 = \frac{2\lambda\gamma^2}{2\lambda'}. \tag{3.49}$$

This trick is applied to ensure that the value of γ'^2 leads us to obtain the same variance of $\boldsymbol{\beta}|\lambda, \gamma^2$. The acceptance probability of λ' is

$$\min\left\{1, \frac{\pi(\lambda')}{\pi(\lambda)} \frac{\left(\prod_{i=1}^{p}\psi_i\right)^{\lambda'-1}}{\left(\prod_{i=1}^{p}\psi_i\right)^{\lambda-1}} \frac{\{\Gamma(\lambda)\}^p}{\{\Gamma(\lambda')\}^p} \frac{\prod_{i=1}^{p}\exp\left(-\frac{\psi_i}{2\gamma'^2}\right)}{\prod_{i=1}^{p}\exp\left(-\frac{\psi_i}{2\gamma^2}\right)} \frac{\left(2\gamma'^2\right)^{p\lambda'}}{(2\gamma^2)^{p\lambda}} \frac{2\lambda}{2\lambda'} \frac{M^2}{M^2} \frac{2\lambda\gamma^2}{2\lambda'\gamma'^2} \frac{\exp\left(-\frac{M}{2\lambda'}\gamma'^{-2}\right)}{\exp\left(-\frac{M}{2\lambda}\gamma^{-2}\right)}\right\} \tag{3.50}$$

It can be seen that some terms can be omitted. For example the terms $\frac{2\lambda\gamma^2}{2\lambda'\gamma'^2}$, and $\frac{\exp\left(-\frac{M}{2\lambda'}\gamma'^{-2}\right)}{\exp\left(-\frac{M}{2\lambda}\gamma^{-2}\right)}$ are constant, because $2\lambda'\gamma'^2 = 2\lambda\gamma^2$ in our updating.

Therefore, we can rewrite Equation 3.50 as follows

$$\min\left\{1, \frac{\pi(\lambda')}{\pi(\lambda)} \frac{\left(\prod_{i=1}^{p}\psi_i\right)^{\lambda'-1}}{\left(\prod_{i=1}^{p}\psi_i\right)^{\lambda-1}} \frac{\{\Gamma(\lambda)\}^p}{\{\Gamma(\lambda')\}^p} \frac{\prod_{i=1}^{p}\exp\left(-\frac{\psi_i}{2\gamma'^2}\right)}{\prod_{i=1}^{p}\exp\left(-\frac{\psi_i}{2\gamma^2}\right)} \frac{(2\gamma^2)^{p\lambda}}{\left(2\gamma'^2\right)^{p\lambda}} \frac{\lambda'}{\lambda}\right\}. \tag{3.51}$$

The acceptance rate is controlled to be between $20\% - 30\%$ using the tuning parameter σ_λ^2.

3.2.5 Full conditional distributions for γ^{-2}

In this section, we will calculate the full conditional distribution for γ^{-2}. The likelihood has no effect on the full conditional distribution of γ^{-2}. Thus, we will derive it from the joint distribution as follows

$$\begin{aligned} f(\gamma^{-2} \mid \beta_i, \psi_i, \lambda, y_i, \boldsymbol{X}) &\propto \left(\gamma^{-2}\right)^{p\lambda} \times \exp\left(-\frac{\sum_{i=1}^{p}\psi_i}{2}\gamma^{-2}\right) \times \left(\gamma^{-2}\right)^{2-1} \times \exp\left(-\frac{M}{2\lambda}\gamma^{-2}\right) \\ &= \left(\gamma^{-2}\right)^{(p\lambda+2)-1} \exp\left\{-\left(\frac{M}{2\lambda} + \frac{1}{2}\sum_{i=1}^{p}\psi_i\right)\gamma^{-2}\right\}. \end{aligned} \tag{3.52}$$

This expression is a gamma distribution with shape parameter $e^* = p\lambda + 2$ and rate parameter $f^* = \frac{M}{2\lambda} + \frac{1}{2}\sum_{i=1}^{p}\psi_i$.

3.3 Approaches to speed up the code and avoid computational problems

In this section, we will explore all specific computational tools which will be implemented to the code to speed up our MCMC updating and avoid computational problems occurring in R.

3.3.1 Approaches to speed up the updating α and β

In section 3.2.2, the full conditional distribution for $\boldsymbol{\phi} = (\alpha, \boldsymbol{\beta})^T$, was calculated and shown to be a multivariate normal distribution. In order to update $\boldsymbol{\phi}$, the Gibbs sampler method is applied to generate samples from the multivariate normal distribution.

Different methods can be used to generate a sample from the multivariate normal distribution, such as using the command '`mvnorm`" in R and using the Cholesky decomposition. A brief explanation of this decomposition is given below.

Cholesky decomposition

The variance-covariance matrix reflects the pairwise linear dependencies among the variables. Like every positive definite symmetric matrix, the covariance matrix can be decomposed into

a product of two matrices, a lower and upper triangular matrix. This decomposition is called the Cholesky decomposition and is used to speed up generating samples from the multivariate normal distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$.

If $\boldsymbol{x} \sim N_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then the Cholesky decomposition method can be applied to generate samples from $\boldsymbol{x}$ as long as $\boldsymbol{\Sigma}$ is positive-definite or positive semi-definite (Gentle, 2009). Let $\boldsymbol{A}$ be a real square matrix such that $\boldsymbol{AA^T} = \boldsymbol{\Sigma}$, where A is a lower triangular matrix with real and positive diagonal elements. Then A is the Cholesky decomposition of $\boldsymbol{\Sigma}$. Let $\boldsymbol{z}$ be a vector of samples generated from the standard normal distribution, then we can sample our vector $\boldsymbol{x}$ via

$$\boldsymbol{x} = \boldsymbol{\mu} + \boldsymbol{Az}. \tag{3.53}$$

In our case, the full conditional distribution for $\boldsymbol{\phi} = (\alpha, \boldsymbol{\beta})^T$ is multivariate normal distribution with mean vector $\overline{\boldsymbol{\phi}} = (\boldsymbol{\Lambda} + \boldsymbol{V}^{-1})^{-1} \boldsymbol{V}^{-1} \hat{\boldsymbol{\phi}}$ and variance-covariance matrix $\boldsymbol{\Sigma} = \text{Var}\left(\boldsymbol{\phi} \mid \hat{\boldsymbol{\phi}}, \boldsymbol{\Lambda}\right) = (\boldsymbol{\Lambda} + \boldsymbol{V}^{-1})^{-1}$. To apply the Cholesky decomposition method in R, one can use the command "`chol`", which returns the upper triangle matrix, but the matrix has to be transposed to obtain the lower triangle matrix, which is required to generate from a multivariate normal distribution. Moreover, one has to be careful with setting the argument of the command "`chol`". If one sets `pivot = FALSE` and the variance-covariance matrix is not positive definite an error message appears. One can set `pivot = TRUE` to avoid such an error. However, if one multiplies the transpose of the output matrix by the output matrix, one would not obtain the original variance-covariance matrix. One can implement three steps to tackle this issue

$$\texttt{Q} = \texttt{chol}\ (\texttt{Sigma}) \tag{3.54}$$

$$\texttt{pivot} = \texttt{attr}\ (\texttt{Q,``pivot''}) \tag{3.55}$$

$$\texttt{Q} = \texttt{Q}[, \texttt{order}(\texttt{pivot})], \tag{3.56}$$

where `Q` is the upper triangle matrix returned from the "`chol`" command, and "`order`" is an R command to rearrange the number ascending or descending order.

In order to update $\boldsymbol{\phi}$ using the Cholesky decomposition method to generate samples from the multivariate normal distribution, the user should follow the following steps

1. Apply the "`chol`" command on the variance-covariance matrix $\boldsymbol{\Sigma}$ with setting `pivot = TRUE`.
2. Calculate the pivot using the command in 3.55.
3. Rearrange the output matrix ($\boldsymbol{Q}$) using the command in 3.56.
4. Generate $p + 1$ independent realisations from a standard normal distribution using the command "`rnorm`" in R.
5. Update $\boldsymbol{\phi}$ using the following equation

$$\boldsymbol{\phi}' = \overline{\boldsymbol{\phi}} + \boldsymbol{Q}^T \boldsymbol{z}, \tag{3.57}$$

where $\boldsymbol{\phi}'$ indicates the updated value of $\boldsymbol{\phi}$, $\overline{\boldsymbol{\phi}}$ represents the vector mean of the full conditional distribution for $\boldsymbol{\phi}$, $\boldsymbol{Q}^T$ is the lower triangle matrix returned from Equation 3.56, $\boldsymbol{z}$ is the vector sampled from standard multivariate normal distribution, and p is the number of variables (SNPs).

The Cholesky decomposition method generates realisations from a multivariate normal distribution faster than using the command "`mvnorm`" in R. Moreover, "`mvnorm`" can cause `R` to crash.

3.3.2 Special cases of the Generalised Inverse Gaussian distribution

The full conditional distribution for ψ_i is calculated in section 3.2.3. The distribution is the generalised inverse Gaussian distribution $GIG\left(m, c, d\right)$ and its density is given by

$$\frac{\left(\frac{c}{d}\right)^{\frac{m}{2}}}{2K_m\left(\sqrt{cd}\right)} x^{m-1} \exp\left\{-\frac{1}{2}\left(cx + \frac{d}{x}\right)\right\}, \tag{3.58}$$

where $m \in \mathbb{R}$ while $c, d \in \mathbb{R}^+$. Additionally, $m = \lambda - \frac{1}{2}, c = \gamma^{-2}$ and $d = \beta_i^2$ (see Equation (3.6)). There are two special cases for the GIG distribution. First, the gamma distribution with

shape m and rate c holds when $m > 0$ and $d = 0$. Second, the inverse gamma distribution with shape m and rate d holds when $m < 0$ and $c = 0$ (Embrechts, 1983).

The value of β_i plays an important role in updating ψ_i, because it determines whether we update ψ_i using the GIG distribution or use its special cases. Therefore, the first step in updating ψ_i is to check whether the value of β_i is zero or not. If the value is not zero, ψ_i will be generated by using the GIG distribution. We use the command "`rgig`" from the "`ghyp`" package in R software (Breymann and Lüthi, 2013) for this purpose. However, if the value of β_i is zero, the special cases will be applied.

Obtaining a small value for ψ_i might affect the variance-covariance matrix of $\boldsymbol{\beta}$, see Equation 3.26. Therefore, some checks are applied to our MCMC to update ψ_i. Initially, we assume that values less than 10^{-5} for β_i^2 are considered to be zero. If $\beta_i^2 \approx 0$ and m is positive $\left(\lambda > \frac{1}{2}\right)$, we use the $Ga\left(\lambda - \frac{1}{2}, \frac{1}{2\gamma^2}\right)$ instead of using the $GIG\left(\lambda - \frac{1}{2}, \frac{1}{2\gamma^2}, \beta_i^2\right)$. Because β_i^2 is not actually zero, the random values generated from the gamma distribution will be tested by comparing $\exp\left(-\frac{\beta_i^2}{\psi_i}\right)$ to the random values generated from uniform distribution $(U \sim (0,1))$ to ensure a correct probabilistic procedure. The value will be accepted if the following condition holds (where U is the uniform distribution)

$$U < \exp\left(-\frac{\beta_i^2}{\psi_i}\right). \tag{3.59}$$

Moreover, if m is negative $\left(\lambda < \frac{1}{2}\right)$ and $\frac{1}{\gamma^2} \approx 0$, we will use the $IGa\left(\lambda - \frac{1}{2}, \beta_i^2\right)$ instead of using $GIG\left(\lambda - \frac{1}{2}, \frac{1}{2\gamma^2}, \beta_i^2\right)$. Because $\frac{1}{\gamma^2}$ is not actually zero, the random values generated from the inverse gamma distribution will be tested by comparing $\exp\left(-\frac{\psi_i}{\gamma^2}\right)$ to the random values generated from uniform distribution $(U \sim (0,1))$ to ensure a correct probabilistic procedure. The value will be accepted if the following condition holds (where U is the uniform distribution)

$$U < \exp\left(-\frac{\psi_i}{\gamma^2}\right). \tag{3.60}$$

In addition, the code will stop working if any generated value of ψ_i has a very small value. Therefore, in such cases we assign any $\psi_i < 10^{-10}$ to be 10^{-10}.

3.4 Using breast cancer to inform the normal gamma (NG) prior

The aim of this section is to use the breast cancer top hits information to inform our Normal-Gamma prior hyperparameters.

3.4.1 Selecting rate parameter κ for λ by minimising sum of squares

Griffin and Brown (2010) suggested in their hierarchal model that $\lambda \sim \text{Exp}(1)$. From Figure 3.1, we showed how the log density of the NG is influenced by the value of λ for a constant value of Var $(\beta \mid \lambda, \gamma^2) = 2\lambda\gamma^2$. The question is what is an appropriate rate parameter for λ in our data analysis? We can exploit the information about the log odds ratios which came from the breast cancer top hits data.

In order to specify the rate parameter κ for λ in the NG prior $(\lambda \sim \text{Exp}(\kappa))$, the following strategy can be implemented. Firstly, we calculated the proportion of the breast cancer top hits (BCTHs) both the discovered SNPs and 1000 yet-be-discovered SNPs below four log OR cutoffs: $\log(1.05)$, $\log(1.08)$, $\log(1.10)$ and $\log(1.2)$. We called these values the "Empirical Probabilities" EP_i. We assumed that all yet-to-be-discovered SNPs had log $\mid$ OR $\mid < \log(1.05)$ since there is little power to detect odds ratios in this range for sample sizes currently used (and none were observed in this range). Next, we applied the Sequential Monte Carlo Method (SMCM) to the NG prior in Equations (3.10) - (3.14) varying κ a cross a suitable range for 1000000 realisations. Then, we calculated the proportion of realisations less than the four log OR cutoffs, and we called these values the "Theoretical Probabilities" TP_i. After that, we calculated $\sum_{i=1}^{4}(TP_i - EP_i)^2$, and we chose κ^*, the rate of λ, that satisfies

$$\kappa^* = \text{argmin}_\kappa \left\{ \sum_{i=1}^{4} (TP_i - EP_i)^2 \right\}. \tag{3.61}$$

The chosen value for κ will be calculated and discussed in Chapter 4.

Chapter 4

Comparing the effectiveness of the methods on simulated data

In the previous chapter we discussed the NG prior in terms of calculating the full conditional distributions for the parameters in the model and explaining the approaches that are used to speed up the code and avoid computational errors. Moreover, we discussed how to exploit the top hits data (see section 1.3) to derive an appropriate value for the exponential rate parameter for λ in the NG prior (see section 3.4.1). The aim of this Chapter is to compare the performance of the NG prior to other multivariate approaches.

One of the common ways of comparing the effectiveness of methods is to apply these methods on simulated data, because in simulated data the user can specify the value of the variables. For example, in our simulated data we know which are the causals SNPs and we can apply the different methods to compare the effectiveness of these methods in identifying the causal SNPs. In this Chapter we will describe the simulated data, how we simulate it, the chosen scenarios of the simulated data and why these scenarios were selected. Moreover, we will discuss the ways of specifying the hyperparameters for Hyper lasso (HL) PiMASS and the NG prior. The aim of this Chapter is to compare the effectiveness of the methods on simulated data using ROC curves (Fawcett, 2006a), to assess the between-dataset variability and to assess how well each method considered copes with the highly correlated nature of the data.

4.1 The simulated data scenarios

In this section we will discuss the scenarios which were selected to simulate our data. There are 8 scenarios that we consider. We used Hapgen2 (Su et al., 2011) to simulate SNPs from a specific region in fine mapping.

Generally when designing experiment, one can use a factorial or fractional factorial design to specify the value of the variables. This is not appropriate here because we already know the effect of the variables. In our case there are three elements that can be varied: odds ratios, sample sizes, and MAF. However, in GWAS the power commonly used to detect the potential causal SNPs is set to be more than 80% for SNP with MAF > 0.05 and this power is a function of odds ratios, sample sizes and MAF. As the values of these elements increase the power also increases. Thus, we effectively have several constraints on the design.

Note that the power is the probability of rejecting the null hypothesis H_0 when the alternative hypothesis H_A is true. In our case can be calculated as follows

$$\text{Power} = P\left(Z \geq C_\alpha \mid Z = \frac{C_\alpha - \beta}{\sqrt{V}}\right), \tag{4.1}$$

where Z is the test statistic, α is the type-I error, C_α is the critical value at α, and β is log odd ratio, and V is the variance that is given $\frac{1}{N \times \text{MAF} \times (1-\text{MAF})}$, where N is the sample size, and MAF is the minor allele frequency (Wakefield, 2009).

Here we are really interested in investigating and quantifying the effect of r^2 between the potential causal SNPs. Therefore, we will consider some scenarios with 2 potential causal SNPs having different LD structure: high correlation between the two potential causal SNPs and no correlation between them. However, the region we simulated from has a finite number of SNPs to choose from. Thus it is difficult to find two SNPs having $r^2 = 0.8$ with a particular MAF. Also r^2 and MAF between the two potential causal SNPs is determined by the haplotype that has been used to generate the genotype data.

4.1.1 Hapgen2

To compare the performance of the methods we need to apply them to simulated data that mimic real data scenarios. Hapgen2 is software developed by Su et al. (2011) to simulate case-control data in fine-mapping studies. This software has the ability to simulate data with multiple causal SNPs. The data are sampled allowing for patterns of linkage disequilibrium (LD). In order to apply Hapgen2, the user has to specify the following inputs

- Specify a file of known haplotypes.
- Specify a legend file for the SNP markers.
- Specify a file containing the fine-scale recombination rate across the region.
- Specify a physical location of SNP, risk allele, heterozygote disease risk odds ratio (OR) and homozygote disease risk $(\mathrm{OR})^2$ assuming an additive model for each causal SNP.
- Specify the number of case and control individuals to sample.
- Specify the beginning and end of the genomic region.
- Specify the name of the output file.

In our simulated data, the European haplotypes of the August 2010 release of the 1000 genomes data was used (Altshuler et al., 2010) as a reference to sample case-control individuals. The result scenarios that simulated by Hapgen2 will be given in the next section.

4.1.2 Scenarios of simulated data

In this section we will discuss the scenarios that are used for simulating case-control data via Hapgen2. There are 8 scenarios simulated to compare the effectiveness of the methods discussed in the Chapter 2 and 3.

Each of the eight scenarios include 2 causal SNPs and these two causal SNPs have the same OR in all scenarios. The OR of the first causal SNP is 1.08 and the OR of the second causal SNP is 1.13. These ORs represent typical effect sizes in fine mapping following

GWAS. However, the scenarios are different in some aspects such as the patterns of linkage disequilibrium (LD) between the causal SNPs, minor allele frequencies of the causal SNPs, the sample size for each scenario, marginal power for each causal SNP and the total number of SNPs. In terms of the patterns of linkage disequilibrium (LD) we specified two patterns: there is high LD and there is almost no LD. If our two causal SNPs are in LD, it can be shown that the LD will not exceed 0.3 for the MAFs we consider. According to VanLiere and Rosenberg (2008), if the following holds:

$$\text{MAF}_1 > \text{MAF}_2 \tag{4.2}$$

the maximum r^2 value between two SNPs is

$$r^2_{\text{max}} = \frac{(1 - \text{MAF}_1) \times \text{MAF}_2}{(1 - \text{MAF}_2) \times \text{MAF}_1}, \tag{4.3}$$

where MAF_1 and MAF_2 are the minor allele frequency of the first causal SNP and the second causal SNP respectively. We calculated r^2_{max} for each two MAFs considered and then selected 2 causal SNPs with r^2 close to this (for the high LD case). In all scenarios the r^2_{max} is between 0.12 and 0.27 (see Tables 4.1 and 4.2). Regarding the minor allele frequency, the first causal SNP is a common SNP in all scenarios with a minor allele frequency of approximately 0.3 and the second causal SNP is a rarer SNP in all scenarios with a minor allele frequency between 0.05 and 0.1. In terms of the sample size of our simulated data, four scenarios have a sample size of 32000 cases and 32000 controls whereas the other 4 scenarios have a sample size of 16000 cases and 16000 controls. In regard to the marginal power for the causal SNPs, the marginal power for the common causal SNP is high (between 0.77 and 0.82) with a sample size of 16000 cases and 16000 controls, whereas the marginal power for the rare causal SNP varied between 0.4 and 0.87 with a sample size of 16000 cases and 16000 controls. However, with a sample size of 32000 cases and 32000 controls, the marginal power for the common causal SNP is high (it is approximately 1), whereas the marginal power for the rare causal SNP varied between 0.9 and approximately 1. The number of SNPs varied from scenario to scenario. The reason behind this is that we remove SNPs which are perfectly correlated with other SNPs and the collinearity between the SNPs varied by scenario. The 8 scenarios are summarised in Tables 4.1 and 4.2.

Scenario	1	2	3	4
r^2 between causal SNPs	0.009	1×10^{-5}	0.13	0.09
r^2_{max} between causal SNPs	0.27	0.14	0.2	0.25
$\frac{r^2}{r^2_{\text{max}}}$	0.03	8×10^{-5}	0.66	0.34
MAF of causal SNP 1	0.28	0.3	0.28	0.31
MAF of causal SNP 2	0.09	0.06	0.07	0.1
Power of causal SNP 1	0.77	0.8	0.77	0.82
Power of causal SNP 2	0.79	0.4	0.55	0.87
Position of causal SNP 1	201689463	201689463	201689463	201689463
Position of causal SNP 2	201803139	201864650	201724114	201793176
Number of SNPs	291	276	281	287

Table 4.1: The values of the specified statistics in the simulated data sets. The scenarios have a total sample size of 32000. The OR of the first causal SNP is 1.08*, whereas the OR of the second causal SNP is* 1.13*.*

4.2 Specifying the parameters for each method

In this section we will discuss the methods used to specify the parameters for Hyper lasso (HL), PiMASS and the hyperparameters for the NG prior. These methods were discussed in detail in Chapters 2 and 3. Here we will identify the values selected for each method.

4.2.1 Specifying the parameters for HL

In section 2.3 we mentioned Hoggart et al. (2008)'s contribution in developing software to implement the HL. To use HL, one has to specify the prior shape and scale. This was explained in section 2.3.3. In our case we specify the shape $\lambda = 0.05$, because it is the smallest value that could be chosen for HL and HL selects a small number of SNPs comparable with the other methods. We calculate the scale γ using Equations (2.27) and (2.29). Therefore the scale for the eight scenarios (see Tables 4.1 and 4.2) was approximately 0.002, 0.004, 0.004, 0.002, 0.004, 0.004, 0.004, 0.004 respectively. Note that α in Equation (2.27) represents the family-wise error rate..

Scenario	5	6	7	8
r^2 between causal SNPs	0.0001	0.002	0.16	0.08
$r^2_{\max}$ between causal SNPs	0.12	0.17	0.22	0.25
$\frac{r^2}{r^2_{\max}}$	0.0005	0.01	0.71	0.32
MAF of causal SNP 1	0.3	0.29	0.27	0.29
MAF of causal SNP 2	0.1	0.05	0.07	0.1
Power of causal SNP 1	0.99	0.99	0.99	0.99
Power of causal SNP 2	0.99	0.9	0.99	0.99
Position of causal SNP 1	201689463	201689463	201689463	201689463
Position of causal SNP 2	201803139	201864650	201724114	201793176
Number of SNPs	281	280	281	280

Table 4.2: The values of the specified statistics in the simulated data sets. The scenarios have a total sample size of 64000. The OR of the first causal SNP is 1.08*, whereas the OR of the second causal SNP is* 1.13*.*

4.2.2 Specifying the parameters for PiMASS

In PiMASS, we used the default values as these seemed to work well. PiMASS was run for 1,000,000 iteration with a 100,000 iteration burn-in.

4.2.3 Specifying the distribution of the hyperparameters for NG

In section 3.4.1 we discussed how to modify the NG prior exploiting the top hits data in order to modify the NG prior for λ. The exponential rate of λ in the NG prior is required to be specified. To achieve this we minimised the sum of squares of the difference in the "Theoretical Probabilities" (TP) and the "Empirical Probabilities" (EP). This was explained in section 3.4.1.

To obtain the optimal value of κ, we varied κ to take values from 0.5 to 200 in increments of 0.5. Then, we calculated the sum of squares for each simulated sample . Figure 4.1 shows plots of the sum of squares in four cases: assuming there are unobserved yet-to-be discovered SNPs totalling 1000, 500, 200 or 100 SNPs (see Section 1.3, and 3.4.1). The figure shows that the value of κ that gives a minimum sum of squares for these cases is: 142.85, 67.1,

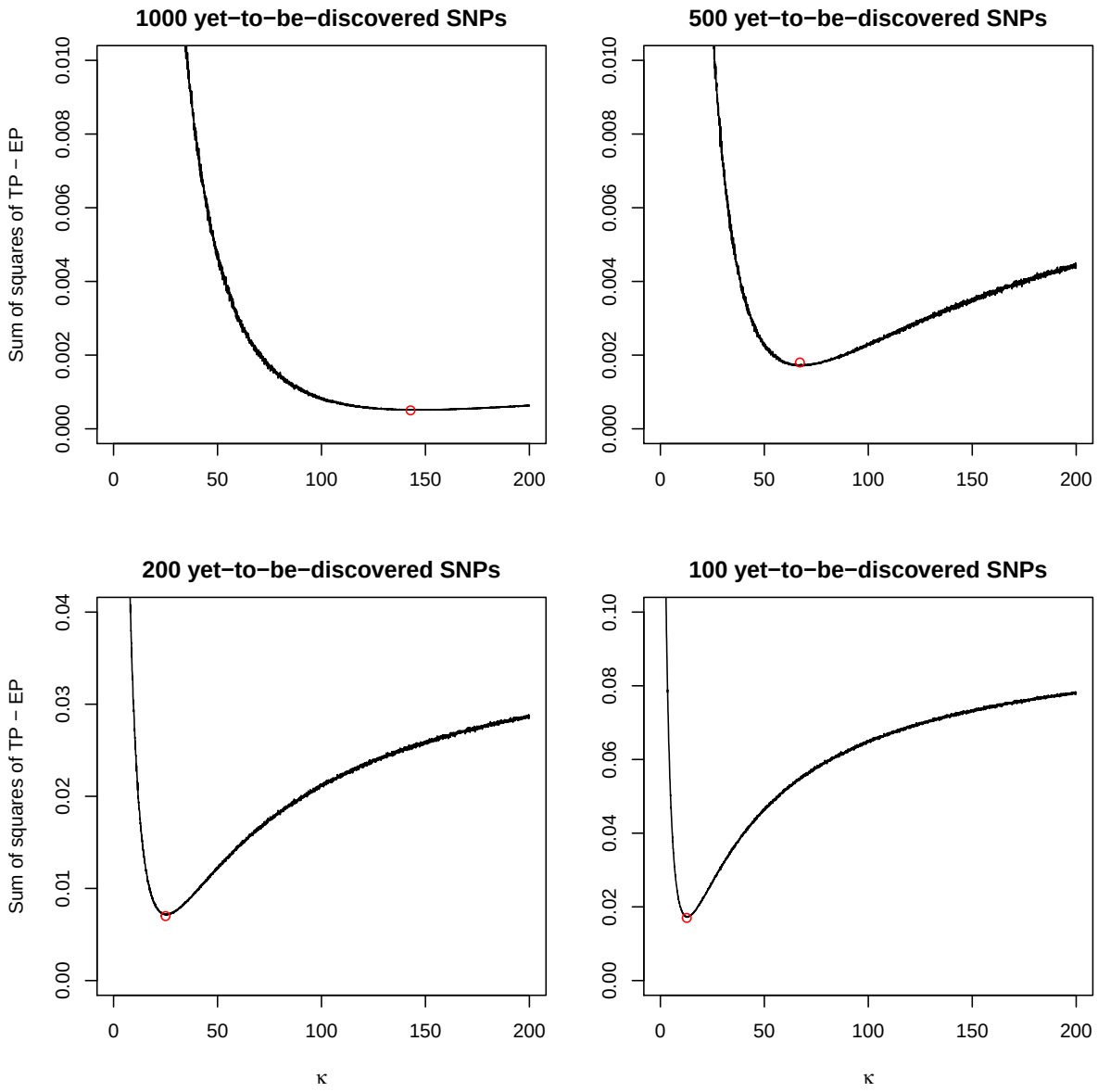


Figure 4.1: The sum of squares of the difference in the "Theoretical Probabilities" (TP) and the "Empirical Probabilities" (EP) versus κ that takes values in $[0.05, 200]$. This plot is for 4 different assumptions regarding the number of SNPs that are yet-to-be discovered: 1000, 500, 200, and 100 yet-to-be discovered SNPs.

25.0, and 12.7 respectively. Here we will assume the first case that there are 1000 unobserved yet-to-be discovered SNPs as suggested in Fachal and Dunning (2015). Moreover, Figure 4.2 shows that 1000 yet-to-be discovered SNPs visually gives the best fit to the empirical ecdf (see Section 3.4.1). Therefore, the optimal κ equals 142.85. Thus, the rate of λ was modified to be 142.85 instead of 0.5 as suggested in Griffin and Brown (2010). The expected value of

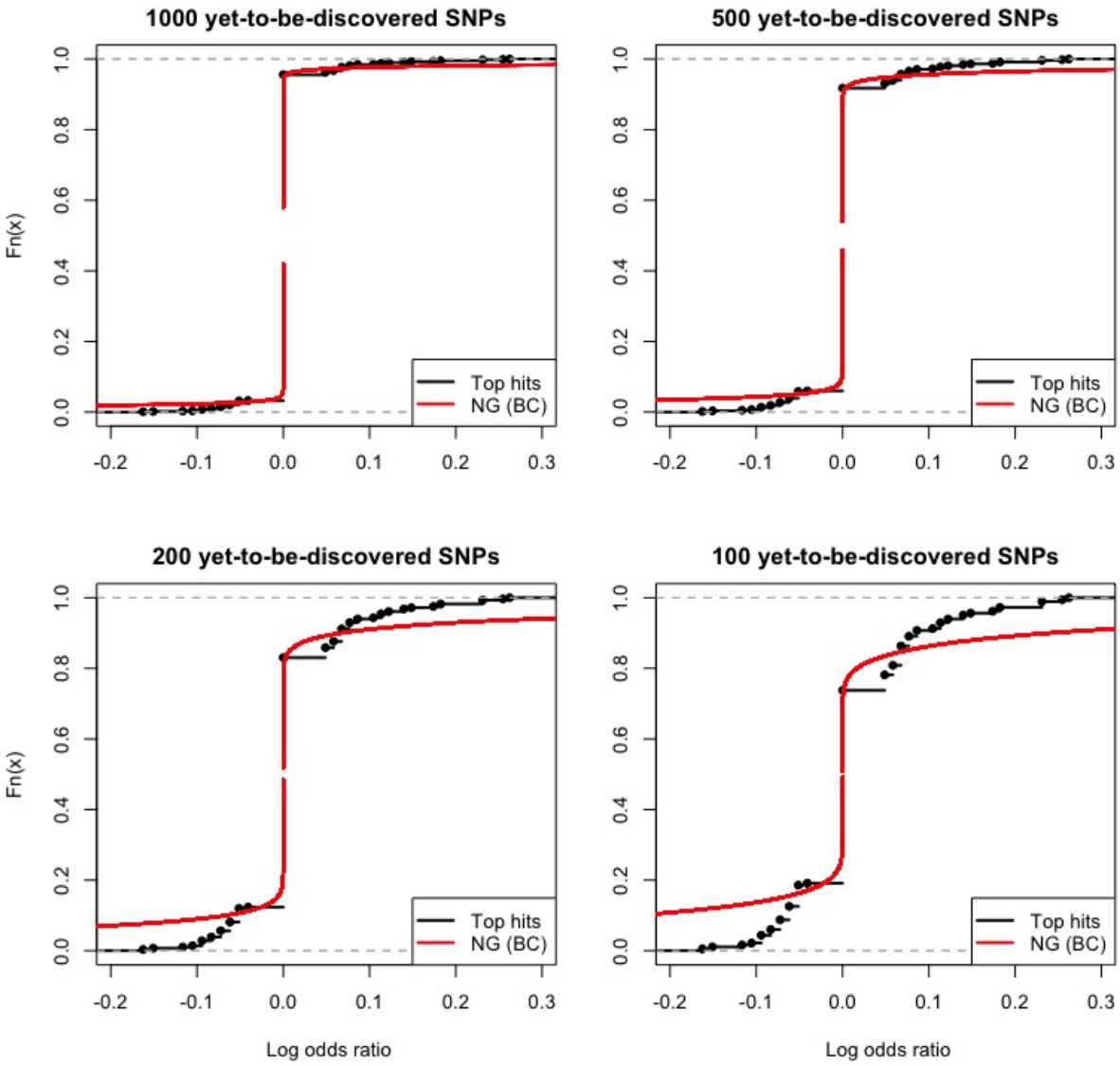


Figure 4.2: ECDF for the top hits data plus some yet-to-be-discovered SNPs (1000, 500, 200, and 100 yet-to-be-discovered SNPs respectively) and ECDF for the values simulated from the sequential Monte Carlo Method for the NG prior with κ minimising the sum squares of the difference in the "Theoretical Probabilities" (TP) and the "Empirical Probabilities" (EP).

λ is therefore $\frac{1}{142.85}$ and Figure 3.1 shows that this leads to a prior with little mass in the tails.

4.3 Comparing the performance of the NG with different κ using ROC curves

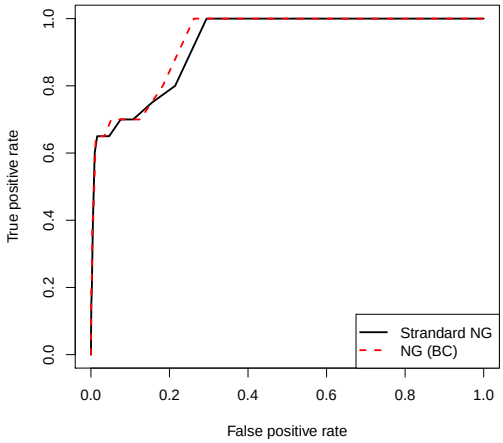
In this section ROC curves are used to compare the performance of the standard NG (with $\kappa = 0.5$) which was used by Griffin and Brown (2010) with the performance of the modified NG (with $\kappa = 142.85$) which was calculated in section 3.4.1. We plotted ROC curves for the 8 scenarios shown in Tables 4.1 and 4.2 with 10 simulated datasets for each scenario. In order to draw the ROC curves, the "ROCR" library (Sing et al., 2005) was used in R. We aggregated the maximum size of credible interval that captured each SNP for each dataset (Fawcett, 2006a) so that we effectively had a single dataset with 20 causal SNPs (See Section 4.6 for more details). To obtain the ROC curves we varied the level of the posterior credible interval calculated as discussed later in Section 4.6. Figure 4.3 shows the 4 scenarios with a sample size of 32000 in Table 4.1, whereas Figure 4.4 shows the 4 scenarios with a sample size of 64000 in Table 4.2.

Figures 4.3 and 4.4 show that generally the performance of the modified NG with $\kappa = 142.85$ is slightly better than the performance of standard NG with $\kappa = 0.5$. Note that the scenarios with sample size of 64000 performs better than the scenarios with sample size of 32000. Therefore, the sample size does have some effect on the performance of the NG prior.

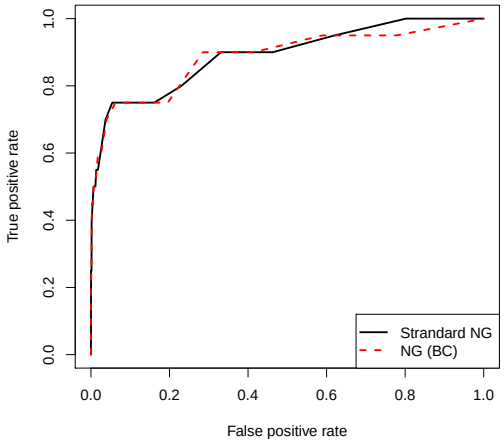
In conclusion, clearly the performance of the modified NG with $\kappa = 142.85$ is better than the performance of standard NG with $\kappa = 0.5$ although the improvement is modest. It is worth noting that the ROC curves are not that sensitive to modifying κ at these large sample sizes. We will therefore only use the NG prior with $\kappa = 142.85$ from now on.

4.4 MCMC trace and acf plot

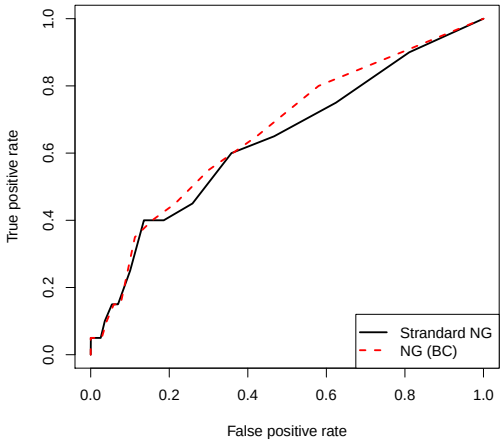
In the section, we will examine the posterior samples obtained using MCMC for the modified NG prior. An important step in MCMC is to examine the posterior samples to check that all parameter posterior distributions converge. To do this, there are two approaches: inspective visualisation (Trace plot and Autocorrelation) (Mengersen et al., 1999) and statistical methods



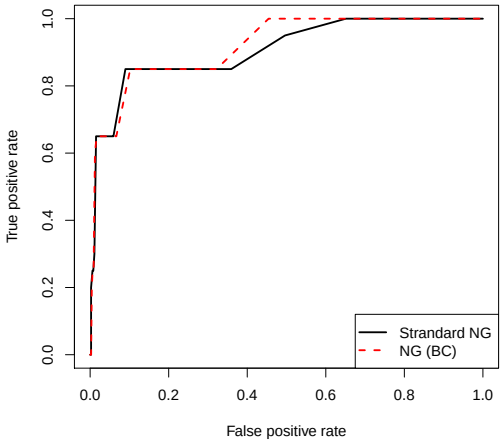
(a) ROC curve of the first scenario with 16000 cases and 16000 controls and $\frac{r^2}{r^2_{\max}} = 0.03$.



(b) ROC curve of the second scenario with 16000 cases and 16000 controls and $\frac{r^2}{r^2_{\max}} = 8 \times 10^{-5}$.

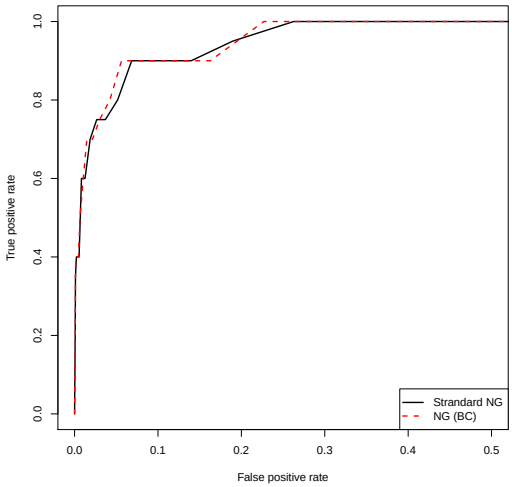


(c) ROC curve of the third scenario with 16000 cases and 16000 controls and $\frac{r^2}{r^2_{\max}} = 0.66$.

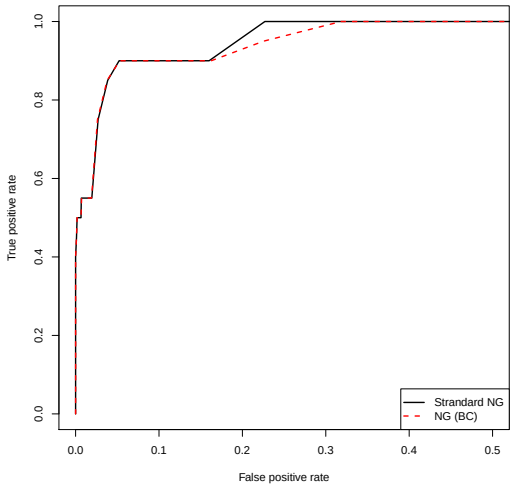


(d) ROC curve of the fourth scenario with 16000 cases and 16000 controls and $\frac{r^2}{r^2_{\max}} = 0.34$.

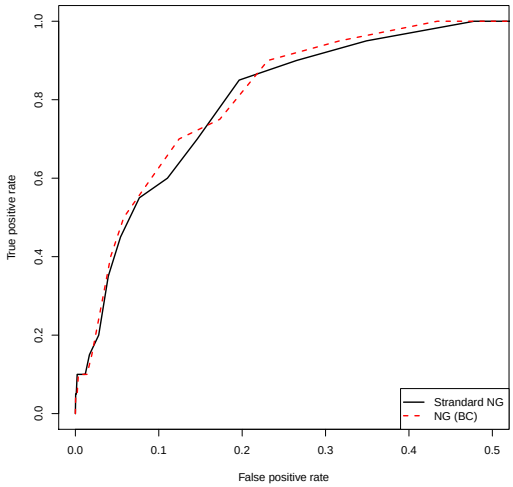
Figure 4.3: ROC curve varying the posterior credible interval of the standard NG prior with $\kappa = 0.5$ and the modified NG prior with $\kappa = 142.85$ NG(BC). Applied to 10 simulated datasets from Hapgen2 with two causal SNPs having odds ratios of 1.08 and 1.13 and having different MAFs (common SNP and rare SNP) described in Table 4.1. Each dataset has 16000 cases and 16000 controls. The MCMC was run for 20,000 iterations with 2,000 burn-in and thinning by 50.



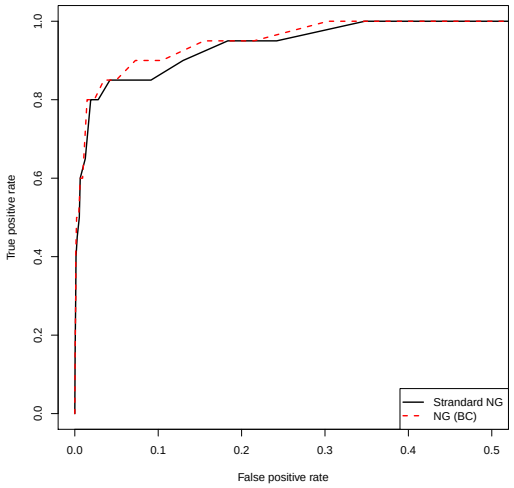
(a) ROC curve of the fifth scenario with 32000 cases and 32000 controls and $\frac{r^2}{r^2_{\max}} = 0.0005$.



(b) ROC curve of the sixth scenario with 32000 cases and 32000 controls and $\frac{r^2}{r^2_{\max}} = 0.01$.



(c) ROC curve of the seventh scenario with 32000 cases and 32000 controls and $\frac{r^2}{r^2_{\max}} = 0.71$.



(d) ROC curve of the eighth scenario with 32000 cases and 32000 controls and $\frac{r^2}{r^2_{\max}} = 0.32$.

Figure 4.4: ROC curve varying the posterior credible interval of the standard NG prior with $\kappa = 0.5$ and the modified NG prior with $\kappa = 142.85$ NG(BC). Applied to 10 simulated datasets from Hapgen2 with two causal SNPs having odds ratios of 1.08 and 1.13 and having different MAFs (common SNP and rare SNP) described in Table 4.2. Each dataset has 32000 cases and 32000 controls. The MCMC was run for 20,000 iterations with 2,000 burn-in and thinning by 50.

($\hat{R}$ method) (Gelman and Rubin, 1992).

Trace plot

A trace plot is a plot of the parameter values at each iteration. We can judge whether the chain reaches equilibrium or not based on the behaviour of the chain. If it does not get stuck at a particular place that means our chain has good mixing (Mengersen et al., 1999).

Figures 4.5(a) and 4.5(c) show the trace plots for the posterior of β and ψ for the common causal SNP from the first dataset, Figures 4.5(b) and 4.5(d) show the trace plots for the posterior of β and ψ for the rare SNP from the first dataset. Figures 4.5(e) and 4.5(f) show the trace plots for the posterior of λ and γ^2 from the first dataset. The MCMC is run for 20,000 iterations with 2,000 iterations burn-in and thinning by 50. Thinning is done because the posterior samples of λ and γ^2 show high levels of autocorrelation (see the next Section). All these figures indicate that the chains were mixing well after thinning so we can rely on the posterior distributions of the parameters that come from MCMC with the modified NG prior with $\kappa = 142.85$.

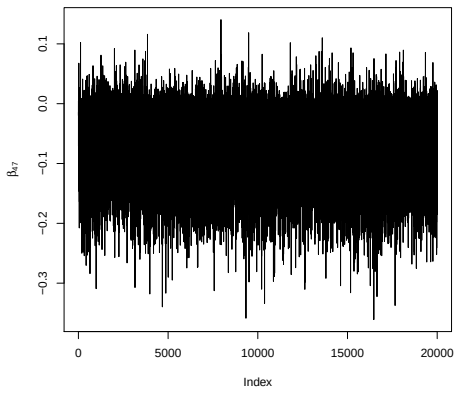
Autocorrelation

Calculating the autocorrelation is another method used to judge whether the MCMC chain has converged or not (Mengersen et al., 1999). In order to do this the lag k autocorrelation ρ_k must be calculated. It is defined as the correlation between each draw of the MCMC chain x_i at lag k and is given by

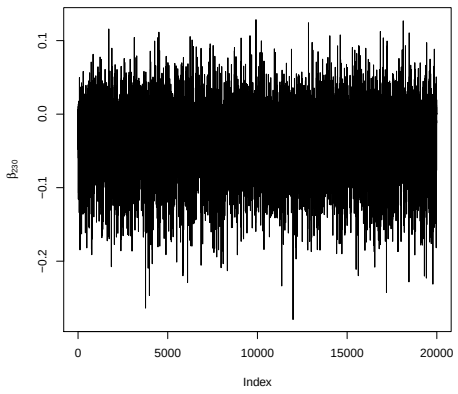
$$\rho_k = \frac{\sum_{i=1}^{n-k} (x_i - \overline{x})(x_{i+k} - \overline{x})}{\sum_{i=1}^{n} (x_i - \overline{x})^2}, \tag{4.4}$$

where $\overline{x}$ represents the mean of the MCMC chain values x_i. An MCMC chain is converged when the autocorrelation between the different draws is approximately 0. Otherwise, the MCMC chain is not converged.

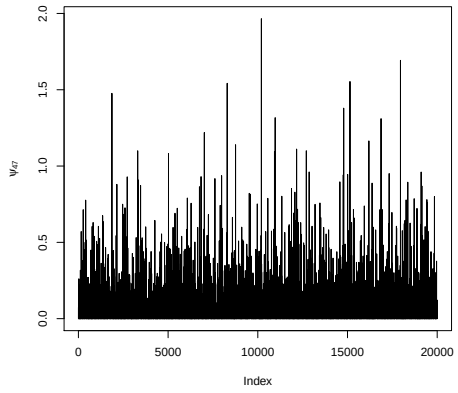
Figures 4.6(a) and 4.6(c) show the ACF plots for the posterior of β and ψ for the common SNP from the first dataset, Figures 4.6(b) and 4.6(d) show the ACF plots for the posterior of β and ψ for the rare SNP from the first dataset, and Figures 4.6(e) and 4.6(f) show the ACF plots for the posterior of λ and γ^2 from the first dataset. The MCMC is run for 20,000



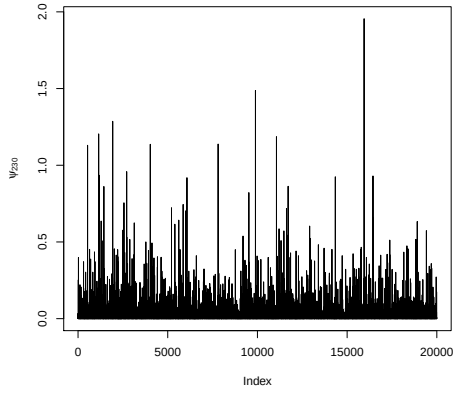
(a) Trace plot of the posterior values of β for the common causal SNP for the first dataset.



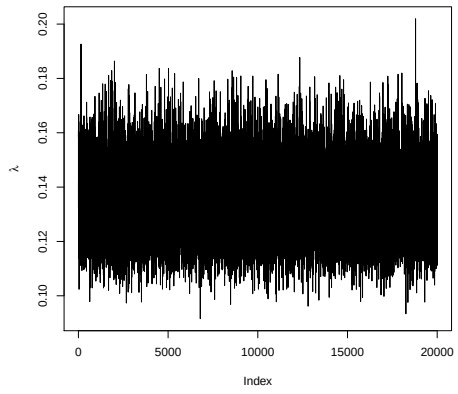
(b) Trace plot of the posterior values of β for the rare causal SNP for the first dataset.



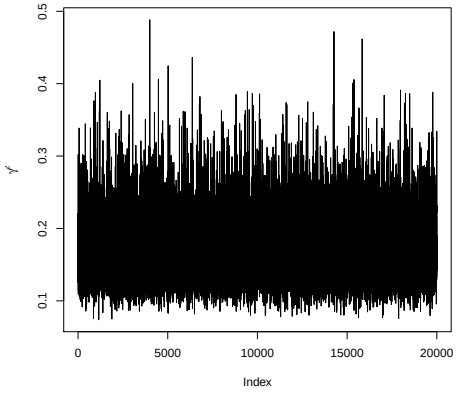
(c) Trace plot of the posterior values of ψ for the common causal SNP for the first dataset.



(d) Trace plot of the posterior values of ψ for the rare causal SNP for the first dataset.



(e) Trace plot of the posterior values of λ for the first dataset.



(f) Trace plot of the posterior values of γ^2 for the first dataset.

Figure 4.5: Trace plots of the posterior values of λ, γ^2, β and ψ for both the common causal SNP and the rare causal SNP using the Normal-Gamma prior. Applied to a simulated dataset from Hapgen2 with two causal SNPs with odds ratios of 1.08 and 1.13 and having different MAF (common SNP and rare SNP). Each dataset has 16000 cases and 16000 controls with 291 SNPs for first scenario. The MCMC was run for 20,000 iterations with 2,000 burn-in and thinning by 50.

iterations with 2,000 iterations burn-in and thinning by 50. All these figures indicate that the autocorrelation between the different draws is approximately 0 once thinning has taken place. Therefore, one can judge that the thinned posterior samples are not too strongly correlated at this level of thinning so we can rely on the posterior distributions of the parameters that come from MCMC with the modified NG prior with $\kappa = 142.85$.

$\widehat{R}$ method

Gelman and Rubin (1992) suggest a multiple sequence diagnostic to assist in the assessment of the convergence of an MCMC chain. First of all, the MCMC must be run more than once (say m times) with different starting values. Let the length of each chain be $2n$. The first half of each chain is removed and the within-chain and between-chain variance of the second half of the chain are calculated. After that the estimated variance of the parameter is calculated as a weighted average of the within-chain and between-chain variance. Finally, the potential scale reduction factor is calculated.

The within-chain variance is the mean of the variance in each chain and it is given by

$$W = \frac{1}{m}\sum_{i=1}^{m} s_i^2, \tag{4.5}$$

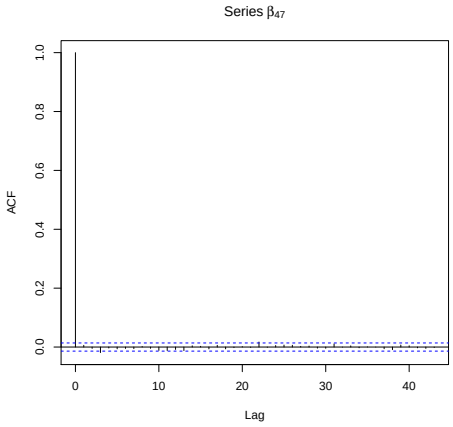
where s_i^2 is the variance of the i^{th} chain. The between-chain variance is given by

$$B = \frac{n}{m-1}\sum_{i=1}^{m}\left(\overline{\theta_i} - \overline{\overline{\theta}}\right)^2, \tag{4.6}$$

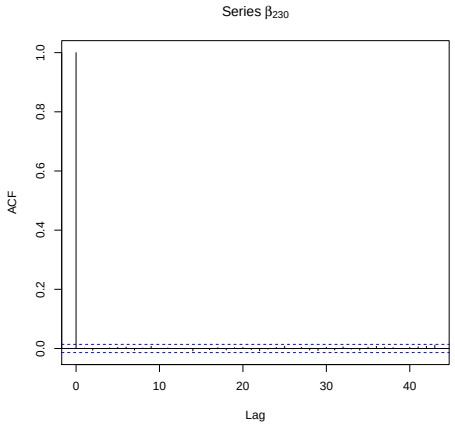
where $\overline{\overline{\theta}}$ is the mean of chain means. The variance of the stationary distribution can be calculated as the weighted average of within-chain variance and between-chain variance. It can be written as follows

$$\widehat{Var}\left(\theta\right) = \left(1 - \frac{1}{n}\right)W + \frac{1}{n}B. \tag{4.7}$$

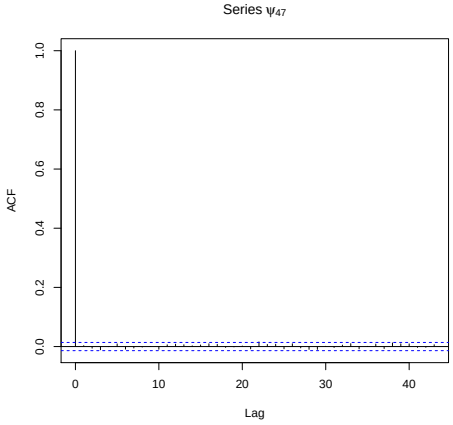
The potential scale reduction factor ($\widehat{R}$) is the square root of the estimated variance of the



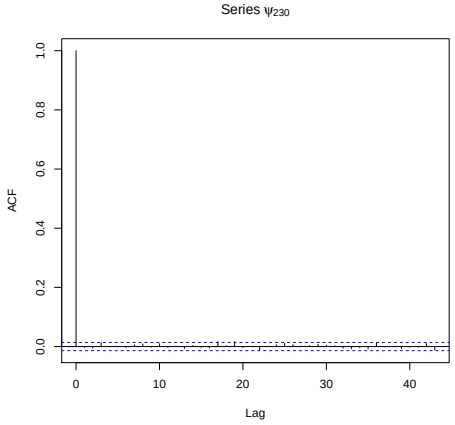
(a) ACF plot of posterior values of β for the common causal SNP for the first dataset.



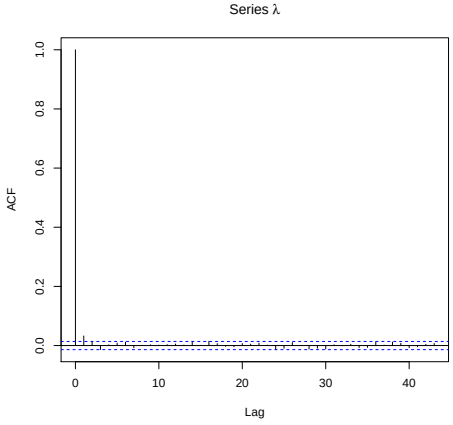
(b) ACF plot of posterior values of β for the rare causal SNP for the first dataset.



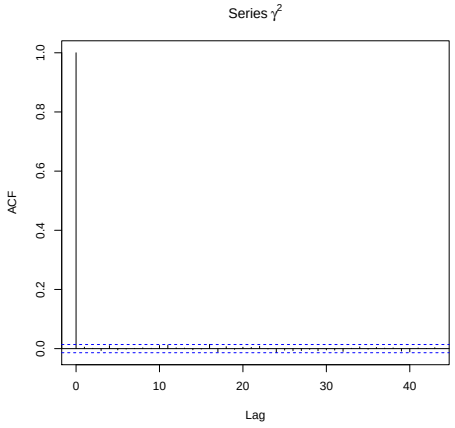
(c) ACF plot of posterior values of ψ for the common causal SNP for the first dataset.



(d) ACF plot of posterior values of ψ for the rare causal SNP for the first dataset.



(e) ACF plot of posterior values of λ for the first dataset.



(f) ACF plot of posterior values of γ^2 for the first dataset.

Figure 4.6: ACF plots of the posterior values of λ, γ^2, β and ψ for both the common causal SNP and the rare causal SNP using the Normal-Gamma prior. Applied to a simulated dataset from Hapgen2 with two causal SNPs with odds ratios of 1.08 and 1.13 and having different MAF (common SNP and rare SNP). Each dataset has 16000 cases and 16000 controls with 291 SNPs for first scenario. The MCMC run for 20,000 iterations with 2,000 burn-in and thinning by 50.

stationary distribution divided by within-chain variance and it can be expressed as follows

$$\widehat{R} = \sqrt{\frac{\widehat{Var}(\theta)}{W}}. \tag{4.8}$$

If the potential scale reduction factor is less than 1.1 then it can be said the chain has converged. In our case, $\widehat{R} < 1.1$ for all parameters checked hence one can judge that all the chains had converged.

4.5 Receiver operating characteristic curves

In this section we will discuss a common way of comparing the performance of different classifying methods, that is receiver operation characteristics (ROC) curves. A ROC curve is a tool for showing the performance of classifier in terms of the true positive rate (sensitivity, TPR), which is the number of positives correctly identified divided by the number of actual positives and the false positive rate (1-specificity, FPR), which is the number of false positives identified divided by the number of negatives. Different thresholds are selected to calculate the previous two quantities and the points generated by varying the threshold are joined to form a curve (actually a step function). Thus, each point on ROC graph refers to a true positive rate and false positive rate at a specific threshold where the y axis represents the true positive rate and the x axis represents the false positive rate. Points at the top left of the plot represent accurate classification.

The area under the ROC curve is a way to quantify the performance of the method. To measure the area under the curve we use the `R` package "pROC" created by Robin et al. (2014). The maximum area under the curve is 1 and the area under the "guess" line is 0.5.

According to Fawcett (2006b) to plot ROC curves for several datasets, there are three methods of averaging the ROC curves: merging averaging, threshold averaging, and vertical averaging. Merging averaging is combining all data sets and calculating the TPR and FPR on the merged data. Vertical averaging fixes the false positive rates and takes the average across datasets of the true positive rates. Threshold averaging fixes the threshold and takes the average across datasets of both the true positive rates and the false positive rates.

In this thesis the ROC curves with merging averaging will be applied to compare between the four methods considered: SLR, HL, PiMASS and the NG prior. Note that the statistics used for each methods are the minimum initial univariate p-value threshold that detects potential causal SNPs in SLR, posterior modes in HL, posterior inclusion probabilities in PiMASS and the maximum credible interval sizes of the posterior in the NG prior that excludes zero.

4.6 Summaries of the NG posterior

In this section, we will discuss some possible posterior summaries using the NG showing which perform well in terms of area under ROC curve. Here we will show three posterior summaries: mean, median and credible intervals for our eight scenarios (see Tables 4.1 and 4.2).

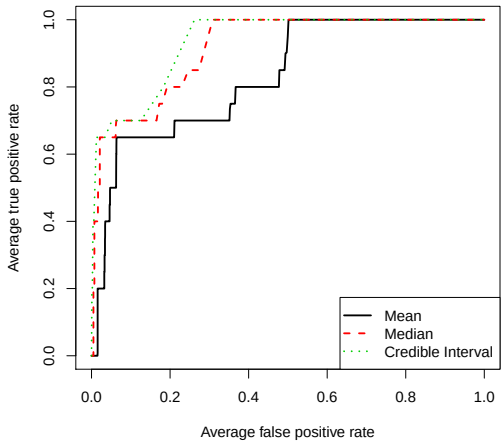
To create the ROC curve, the `R` package ``ROCR'' (Sing et al., 2005) was applied to create the ROC curve for both posterior mean and posterior median. The method used is "merging" the posterior summaries of 10 datasets into one vector (Fawcett, 2006a). However, the ROC curves for varying the credible intervals were created via a code that was programmed using `R`. To create the ROC curve by varying the credible interval, the interval level is varied as a sequence taking values $0\%, 1\%, \ldots, 100\%$. Then the credible interval is calculated at each level. If the credible interval includes 0, the SNP is recorded as a non-causal SNP. Then, each SNP is represented by the maximum credible interval size that detects it, hence each SNP would take a value between 0 and 1. For example, the maximum credible interval size that detects the common causal SNP in first dataset of the first scenario is 0.6. It can be seen in Table 4.8 that the common causal SNPs is not detected at 70%, whereas it is detected at 60%.Finally, the ROC curves were plotted by merging the posterior credible interval sizes for 10 datasets into one vector and the false positive rates (FPR) and the true positive rates (TPR) were calculated as the threshold varied. We call this approach " Varying the credible interval". Alternatively, one can calcualte

$$CI = 1 - 2 \min \left\{ Pr\left(\beta \mid \hat{\beta}, V\right) > 0, Pr\left(\beta \mid \hat{\beta}, V\right) < 0 \right\}. \tag{4.9}$$

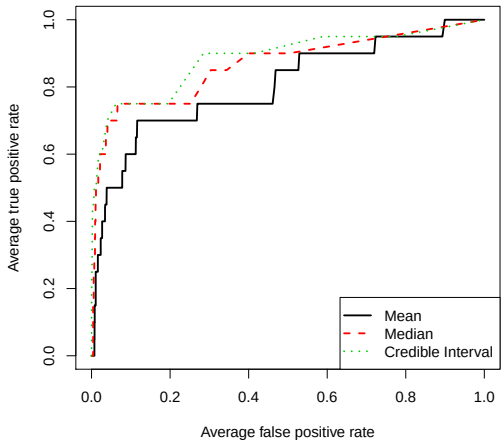
Figures 4.7 and 4.8 show the ROC curves for the posterior mean, median and varying the credible interval for the NG prior with $\kappa = 142.85$. It can be seen that there are quite big differences in the performance of the three posterior summaries for the modified NG prior for all eight scenarios. In Figures 4.9 and 4.10 we zoom in on the plot to consider only the false positive rate from 0 to 0.5. It can be seen that the performance of the posterior credible interval is clearly better than the performance of the other posterior summaries (median and mean), because it nearly always detects more causal SNPs for a given number of non-causal SNPs selected. Moreover, in each scenario the curve of the posterior mean is the worst curve and it has the lowest area under the curve. Therefore, we will only use the credible interval approach from now on. The posterior mean and median could be used for ranking SNPs but if a hard decision is needed then the credible interval approach permits this and so is also the statistic that is the most interpretable.

4.7 Comparing the performance of NG, HL, PiMASS, and SLR

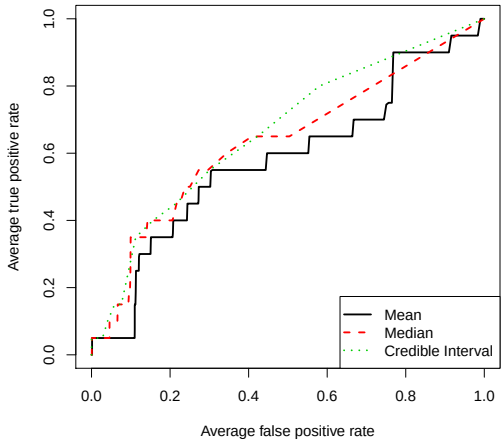
In this section, we will compare four methods: Hyper lasso, PiMASS, NG, and Sequential Logistic Regression. Hyper lasso (HL) was run with shape = 0.05 and scale calculated based on the number of SNPs in a particular scenario. The scale was approximately 0.002, 0.004, 0.004, 0.002, 0.004, 0.004, 0.004, and 0.004 for scenario1-scenario8 respectively. Note that a family-wise error rate (FWER) of 0.05 was used (see Section 4.2.1). PiMASS was run using the default values of the parameters and run for 1,000,000 iterations with 100,000 burn-in. The posterior credible interval of the NG prior with $\kappa = 142.85$ was used where the MCMC was run for 20,000 iterations with 2,000 burn-in and thinned by 50. We also used Sequential Logistic Regression (SLR) varying the initial univariate threshold from 1×10^{-5} to 0.99, and applying forward stepwise logistic regression (see Section 2.2). `R` package "ROCR" (Sing et al., 2005) was applied to create the ROC curves for both the effect sizes from HL and the posterior inclusion probability (PIP) from PiMASS. We wrote `R` code to plot the ROC curves for the posterior credible interval for the NG prior as demonstrated in Section 4.6 and for SLR



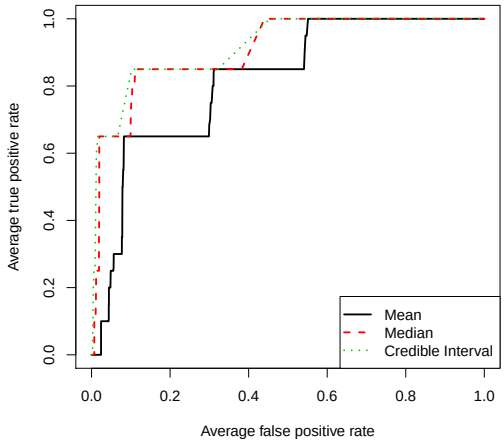
(a) ROC curve of the first scenario with 16000 cases and 16000 controls and $\frac{r^2}{r^2_{\max}} = 0.03$.



(b) ROC curve of the second scenario with 16000 cases and 16000 controls and $\frac{r^2}{r^2_{\max}} = 8 \times 10^{-5}$.

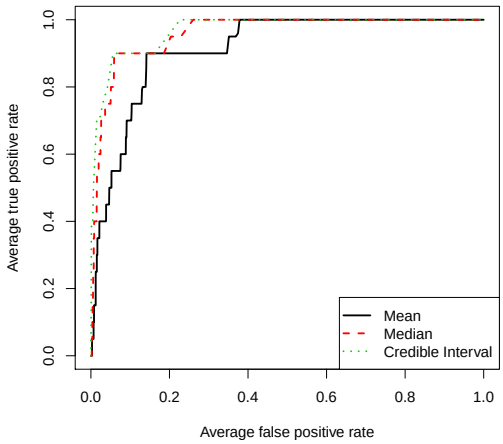


(c) ROC curve of the third scenario with 16000 cases and 16000 controls and $\frac{r^2}{r^2_{\max}} = 0.66$.

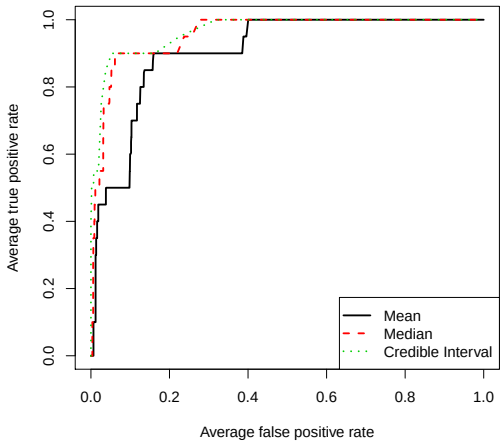


(d) ROC curve of the fourth scenario with 16000 cases and 16000 controls and $\frac{r^2}{r^2_{\max}} = 0.34$.

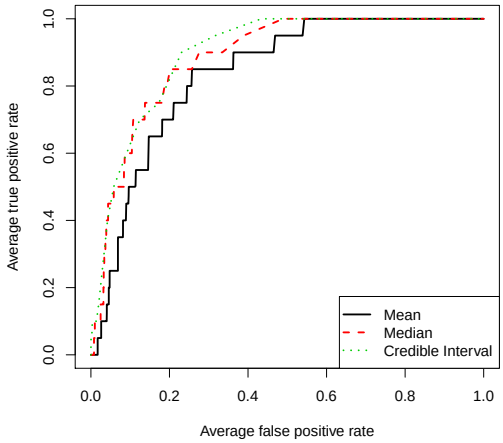
Figure 4.7: ROC curve for the posterior mean, median and varying the credible interval for the Normal-Gamma prior with $\kappa = 142.85$. The methods are applied to 10 simulated datasets from Hapgen2 with two causal SNPs with odds ratios of 1.08 and 1.13 and different MAFs (see Table 4.1). Each dataset has 16000 cases and 16000 controls. The MCMC was run for 20,000 iterations with 2,000 burn-in and thinning by 50.



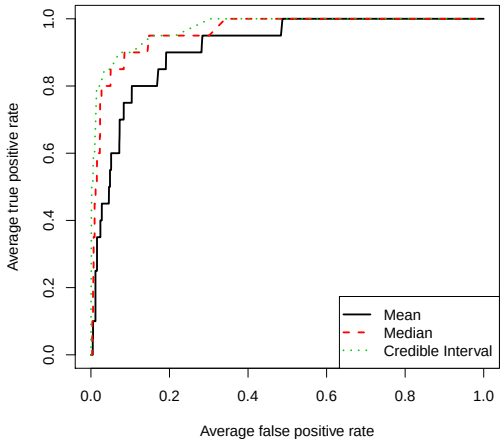
(a) ROC curve of the first scenario with 32000 cases and 32000 controls and $\frac{r^2}{r^2_{\max}} = 0.0005$.



(b) ROC curve of the second scenario with 32000 cases and 32000 controls and $\frac{r^2}{r^2_{\max}} = 0.01$.

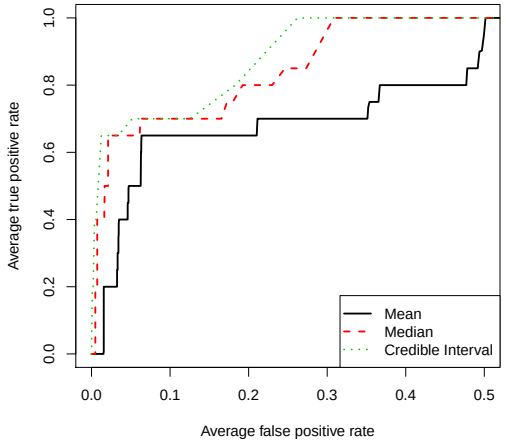


(c) ROC curve of the third scenario with 32000 cases and 32000 controls and $\frac{r^2}{r^2_{\max}} = 0.71$.

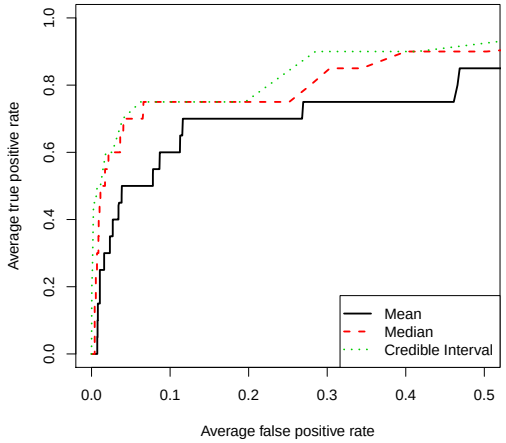


(d) ROC curve of the fourth scenario with 32000 cases and 32000 controls and $\frac{r^2}{r^2_{\max}} = 0.32$.

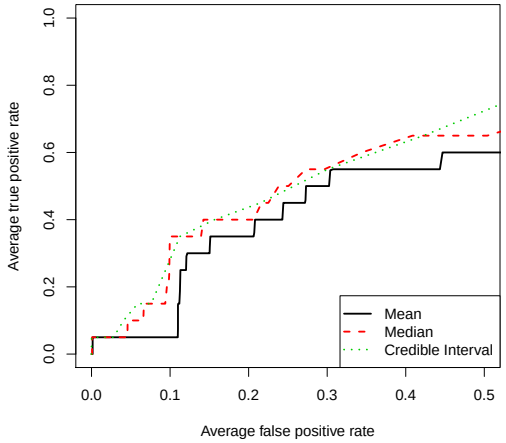
Figure 4.8: ROC curve for the posterior mean, median and varying the credible interval for the Normal-Gamma prior with $\kappa = 142.85$. The methods are applied to 10 simulated datasets from Hapgen2 with two causal SNPs with odds ratios of 1.08 and 1.13 and different MAFs (see Table 4.2). Each dataset has 32000 cases and 32000 controls. The MCMC was run for 20,000 iterations with 2,000 burn-in and thinning by 50.



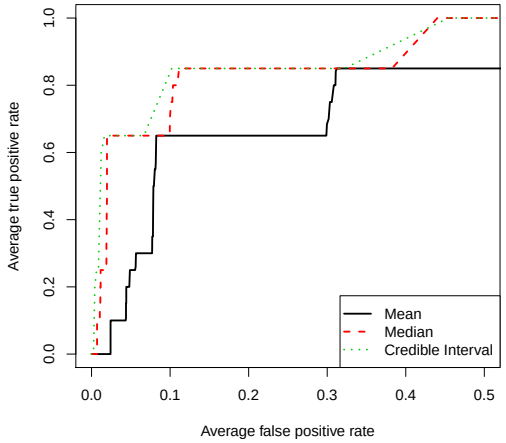
(a) Zooming ROC curve of the first scenario with 16000 cases and 16000 controls and $\frac{r^2}{r^2_{\max}} = 0.03$.



(b) Zooming ROC curve of the second scenario with 16000 cases and 16000 controls and $\frac{r^2}{r^2_{\max}} = 8 \times 10^{-5}$.

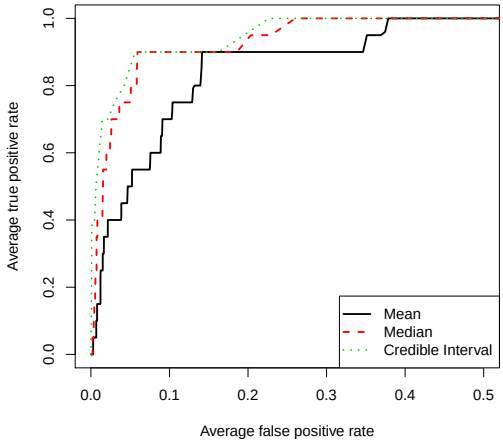


(c) Zooming ROC curve of the third scenario with 16000 cases and 16000 controls and $\frac{r^2}{r^2_{\max}} = 0.66$.

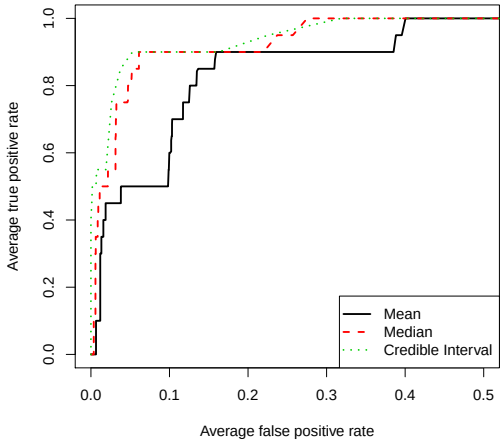


(d) Zooming ROC curve of the fourth scenario with 16000 cases and 16000 controls and $\frac{r^2}{r^2_{\max}} = 0.34$.

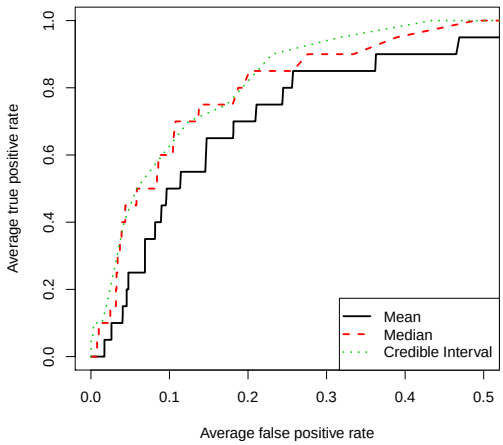
Figure 4.9: ROC curve for the posterior mean, median and varying the credible interval for the Normal-Gamma prior with $\kappa = 142.85$ for FPR < 0.5. The methods are applied to 10 simulated datasets from Hapgen2 with two causal SNPs with odds ratios of 1.08 and 1.13 and different MAFs (see Table 4.1). Each dataset has 16000 cases and 16000 controls. The MCMC was run for 20,000 iterations with 2,000 burn-in and thinning by 50.



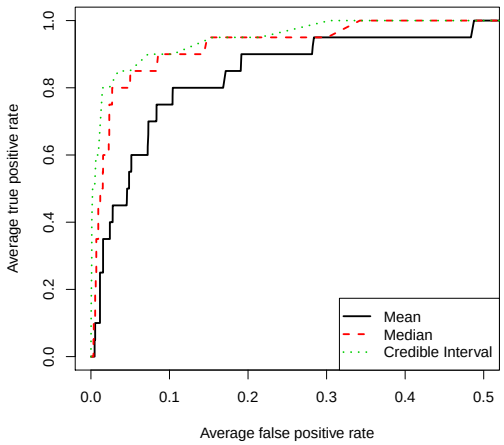
(a) Zooming ROC curve of the first scenario with 32000 cases and 32000 controls and $\frac{r^2}{r^2_{\max}} = 0.0005$.



(b) Zooming ROC curve of the second scenario with 32000 cases and 32000 controls and $\frac{r^2}{r^2_{\max}} = 0.01$.



(c) Zooming ROC curve of the third scenario with 32000 cases and 32000 controls and $\frac{r^2}{r^2_{\max}} = 0.71$.



(d) Zooming ROC curve of the fourth scenario with 32000 cases and 32000 controls and $\frac{r^2}{r^2_{\max}} = 0.32$.

Figure 4.10: ROC curve for the posterior mean, median and varying the credible interval for the Normal-Gamma prior with $\kappa = 142.85$ for FPR < 0.5. The methods are applied to 10 simulated datasets from Hapgen2 with two causal SNPs with odds ratios of 1.08 and 1.13 and different MAFs (see Table 4.2). Each dataset has 32000 cases and 32000 controls. The MCMC was run for 20,000 iterations with 2,000 burn-in and thinning by 50.

by calculating the true positive rate (TPR) and false positive rate (FPR) of the selected SNPs from the SLR at each threshold (see Section 2.2). Note that for SLR, the ROC space contains a set of points where each point refers to the average TPR of the 10 datasets and the average FPR of the 10 datasets for a particular univariate threshold. The methods are applied to 10 datasets for 8 scenarios (see Table 4.1 and 4.2).

Scenarios	HL	NG	PiMASS
1	0.92	0.93	0.91
2	0.65	0.89	0.84
3	0.72	0.67	0.65
4	0.82	0.92	0.79
5	0.9	0.97	0.99
6	0.87	0.97	0.99
7	0.83	0.9	0.85
8	0.98	0.97	0.99

Table 4.3: Area under the curve for HL, NG and PiMASS for all eight scenarios

Figures 4.11 and 4.13 $(\text{FPR} < 0.5)$ show the ROC curves for 3 methods (HL, NG, and PiMASS) and points for SLR using the scenarios with 32000 sample size. They indicate that the performance of HL varies from scenario to scenario based on the level of LD and the marginal power. Note that Figures 4.11(a) and 4.13(a) indicate that the performance of HL in the first scenario with a low level of LD and high marginal power for both causal SNPs is the best HL performance among other scenarios, but is not the best among other approaches (PiMASS, NG and SLR). Figures 4.11(b) and 4.13(b) indicate that the performance of HL in the second scenario with a low level of LD and low marginal power to detect the rare causal SNP is not only the worst performance for HL among other scenarios, but also among all other three methods (PiMASS, NG and SLR). This difference in performance appears to be because of the low marginal power for the rarer causal SNPs in Figure 4.11(b) relative to that

in Figure 4.11(a) (see Table 4.1). Additionally, Figures 4.11 and 4.13 show the performance of the posterior credible interval for the NG prior is competitive with that of the other methods in all the four scenarios except for the third scenario where the performance of HL is better in terms of the area under the curve (Table 4.3 indicates that the area under curve for HL is approximately 0.72 whereas in the posterior credible interval for the NG prior it is approximately 0.67). This indicates that the posterior credible interval for the NG prior struggles with the high level of LD. In addition, the performance of PiMASS is clearly affected by the level of LD. Note its performance in the third and fourth scenarios (high level of LD and moderate level of LD respectively) compared to its performance in the first and second scenario (low level of LD) (see Figures 4.11 and 4.13). The area under the curve for PiMASS in all four scenarios is approximately 0.91, 0.84, 65, and 0.79 respectively (see Table 4.3). Moreover, in SLR, it can be seen that the points in the first and second scenarios are recorded with higher true positive rates than the third and fourth scenarios. The level of LD could also be the reason behind this. In all scenarios, it can be seen that all points from SLR are generally located below at least one of the three curves (HL, NG, and PiMASS) with the notable exception of the point in the first scenario that refers to the initial univariate threshold of 1×10^{-3} which has the highest true positive rate with a very small false positive rate.

Figures 4.12 and 4.14 show the ROC curves for 3 methods (HL, NG, and PiMASS) and points for SLR using the scenarios with 64000 sample size. Generally they indicate that the performance of all methods with a larger sample size is improved which is to be expected. Interestingly, the performance of PiMASS is improved dramatically and its performance at this sample size is sometimes now better than the other approaches. In the seventh scenario the performance of the posterior credible interval for the NG prior is the best curve according to the area under the curve (Table 4.3 indicates that the area under curve for HL is approximately 0.83, the area under curve for PiMASS is approximately 0.85 whereas in the posterior credible interval for the NG prior it is approximately 0.9 although at low FPRs SLR and HL capture more causal SNPs). Again this might be due to the influence of the high level of LD.

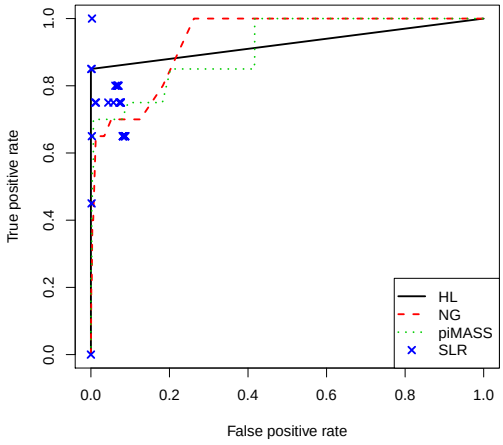
In conclusion, both the level of LD and the marginal power seem to affect the performance of the approaches in fine mapping studies in a complicated way. All four methods struggled with the high level of LD. As expected, increasing the sample size improves the performance

as we have seen in Figures 4.12 and 4.14. For example, the area under the curve of PiMASS in the first scenario was approximately 0.91 compared to approximately 0.99 in the fifth scenario (see Table 4.3).

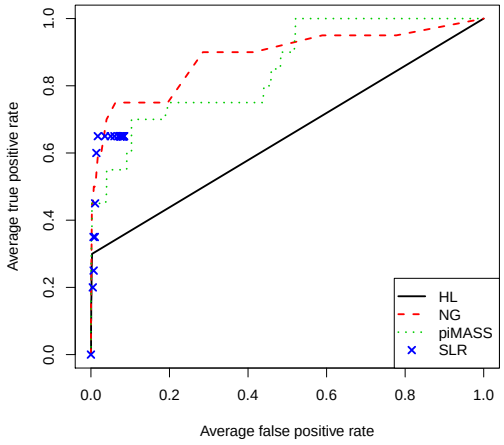
4.8 Between-dataset variability in performance of the four methods

Figures 4.11 - 4.14 indicate the performance of combining the effect sizes from HL, posterior inclusion probabilities (PIP) from PiMASS and the credible interval of the posterior NG for 10 datasets and indicate the average performance measured by TPR and FPR across 10 datasets for SLR but it is also of interest to examine the between-dataset variability. Throughout this section we discuss between-dataset variability in the false positive rate (FPR) at the fixed true positive rates (TPR) of 0.5 and 1 for HL, PiMASS and the NG prior. For a single dataset the FPR is not necessarily unique at a given TPR. Therefore, we use the lowest FPR at the given TPR. This approach was applied to the four scenarios with a total sample size of 32000 (see Table 4.1).

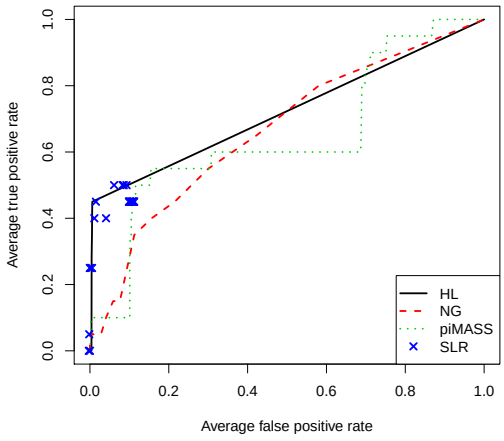
Tables 4.4 - 4.5 show $1000\times$ FPR for each dataset for all scenarios mentioned in Table 4.1 using the NG Credible interval sizes, HL, PiMASS, and SLR all at TPR $= 0.5$ and TPR $= 1$. An NA indicates the FPR at the corresponding TPR is not available. This happens when HL does not select either causal SNP and when SLR either none or both causal SNPs are captured. Within each method the left (right) column corresponds to TPR = 0.5 (TPR = 1). In general, Tables 4.4 - 4.5 indicate that the credible interval of the posterior NG is the most variable approach. Moreover, as the level of LD increases (Scenario 3 and 4) the FPR increases at the given TPR and this explains the reduction in the area under the curve for the NG approach (see Figures 4.11(c) and 4.13(c)). For the fourth scenario, noticed that FPR at TPR = 0.5 for the NG in the third scenario for first and tenth datasets are not available (NA). This means both causal SNPs have exactly the same CI. In addition, it can be seen that FPR for HL in the second and fourth scenario for some datasets are not available (NA) because they fail to detect either causal SNP. Therefore the performance of HL was the worst in these



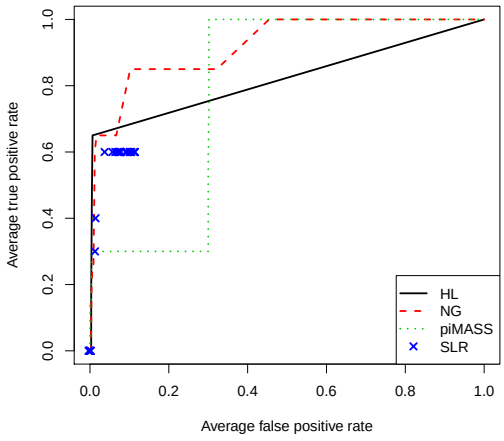
(a) ROC curve of the first scenario with 16000 cases and 16000 controls and $\frac{r^2}{r^2_{\max}} = 0.03$.



(b) ROC curve of the second scenario with 16000 cases and 16000 controls and $\frac{r^2}{r^2_{\max}} = 8 \times 10^{-5}$.

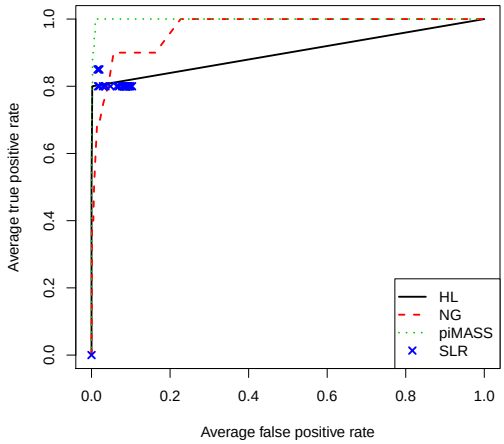


(c) ROC curve of the third scenario with 16000 cases and 16000 controls and $\frac{r^2}{r^2_{\max}} = 0.66$.

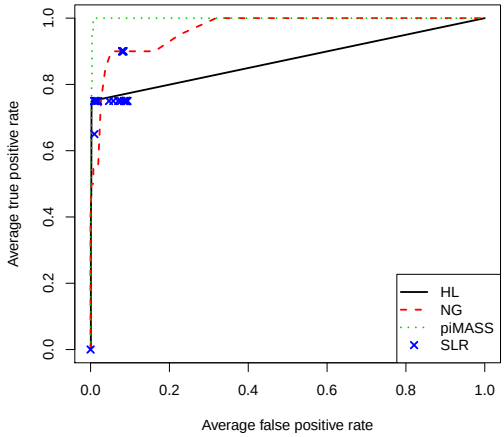


(d) ROC curve of the fourth scenario with 16000 cases and 16000 controls and $\frac{r^2}{r^2_{\max}} = 0.34$.

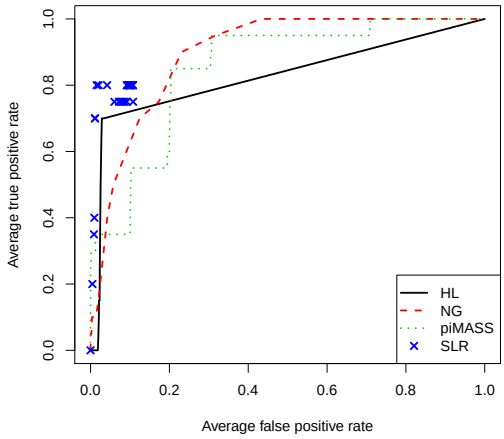
Figure 4.11: ROC curves for the stepwise logistic regression (SLR), PiMASS, Hyper lasso (HL) and the posterior credible interval using Normal-Gamma prior (NG) for an asymptotic normal likelihood. The methods are applied to 10 simulated dataset from Hapgen2 with two causal SNPs with odds ratios of 1.08 and 1.13 and having different MAFs (see Table 4.1). Each dataset has 16000 cases and 16000 controls. The NG MCMC was run for 20,000 iterations with 2,000 burn-in and thinning by 50. PiMASS is used with the default parameters. Hyper lasso is implemented with shape equal to 0.05 and scale equal to 0.002, 0.004, 0.004, and 0.002 respectively.



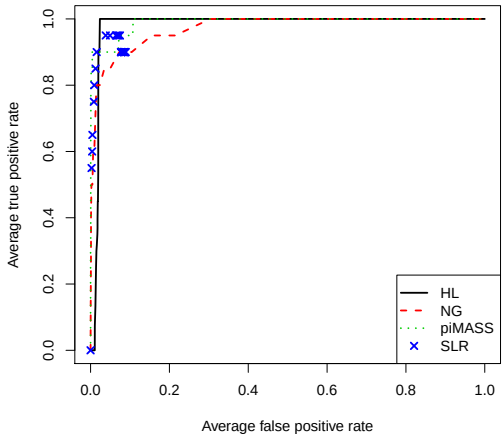
(a) ROC curve of the fifth scenario with 32000 cases and 32000 controls and $\frac{r^2}{r^2_{\max}} = 0.0005$.



(b) ROC curve of the sixth scenario with 32000 cases and 32000 controls and $\frac{r^2}{r^2_{\max}} = 0.01$.

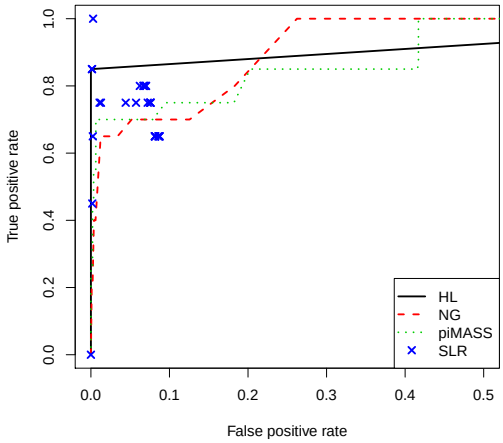


(c) ROC curve of the seventh scenario with 32000 cases and 32000 controls and $\frac{r^2}{r^2_{\max}} = 0.71$.

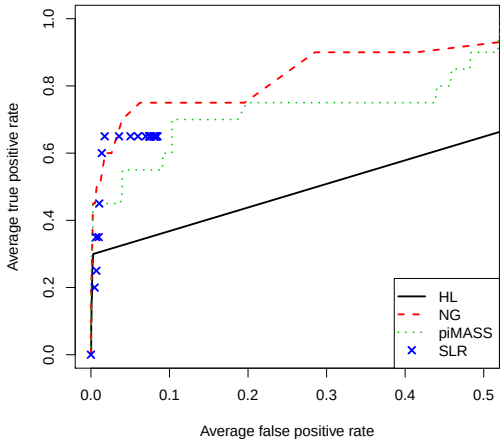


(d) ROC curve of the eighth scenario with 32000 cases and 32000 controls and $\frac{r^2}{r^2_{\max}} = 0.32$.

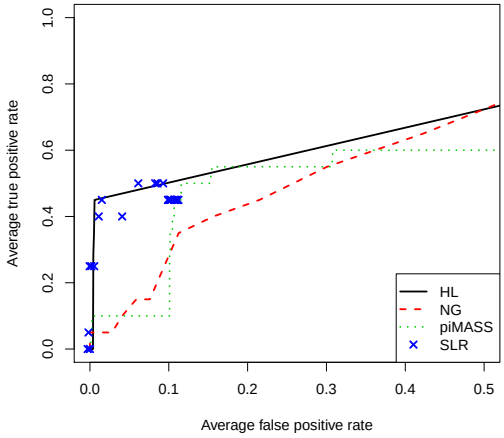
Figure 4.12: ROC curves for the stepwise logistic regression (SLR), PiMASS, Hyper lasso (HL) and the posterior credible interval using Normal-Gamma prior (NG) for an asymptotic normal likelihood. The methods are applied to 10 simulated dataset from Hapgen2 with two causal SNPs with odds ratios of 1.08 and 1.13 and having different MAFs (see Table 4.2). Each dataset has 32000 cases and 32000 controls. The NG MCMC was run for 20,000 iterations with 2,000 burn-in and thinning by 50. PiMASS is used with the default parameters. HyperLASSO is implemented with shape equals 0.05 and scale equals 0.004.



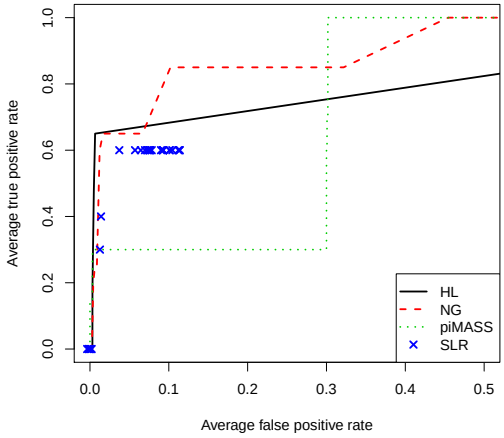
(a) Zooming ROC curve of the first scenario with 16000 cases and 16000 controls and $\frac{r^2}{r^2_{\max}} = 0.03$.



(b) Zooming ROC curve of the second scenario with 16000 cases and 16000 controls and $\frac{r^2}{r^2_{\max}} = 8 \times 10^{-5}$.

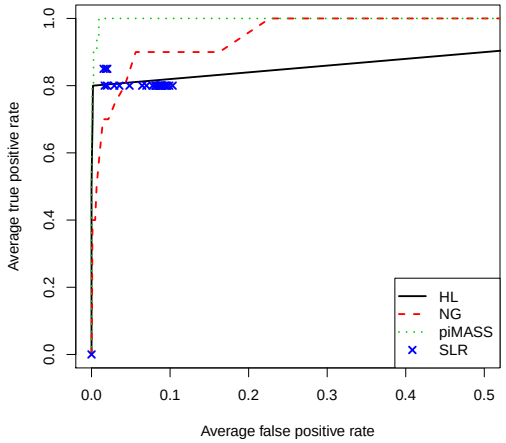


(c) Zooming ROC curve of the third scenario with 16000 cases and 16000 controls and $\frac{r^2}{r^2_{\max}} = 0.66$.

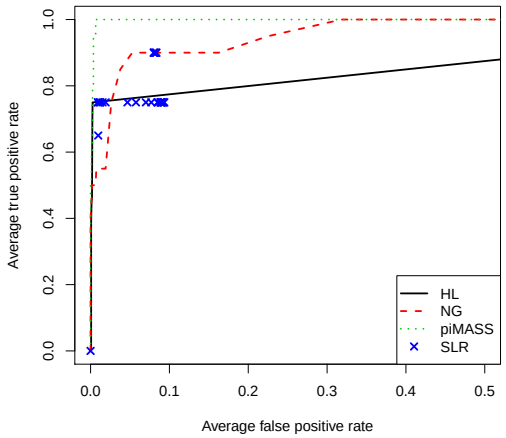


(d) Zooming ROC curve of the fourth scenario with 16000 cases and 16000 controls and $\frac{r^2}{r^2_{\max}} = 0.34$.

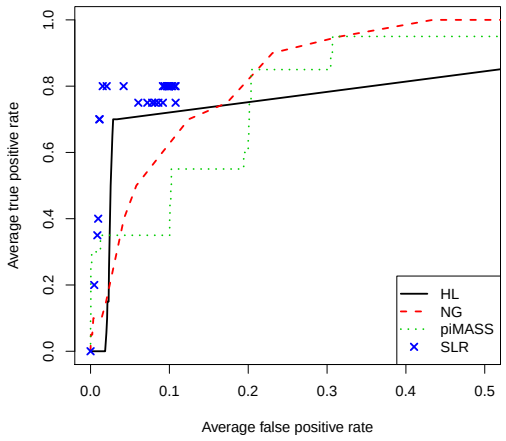
Figure 4.13: ROC curves $(FPR \leq 0.5)$ for the stepwise logistic regression (SLR), PiMASS, Hyper lasso (HL) and the posterior credible interval using Normal-Gamma prior (NG) for an asymptotic normal likelihood. The methods are applied to 10 simulated dataset from Hapgen2 with two causal SNPs with odds ratios of 1.08 and 1.13 and having different MAFs (see Table 4.1). Each dataset has 16000 cases and 16000 controls. The NG MCMC was run for 20,000 iterations with 2,000 burn-in and thinning by 50. PiMASS is used with the default parameters. Hyper lasso is implemented with shape equal to 0.05 and scale equal to 0.002, 0.004, 0.004, and 0.002 respectively.



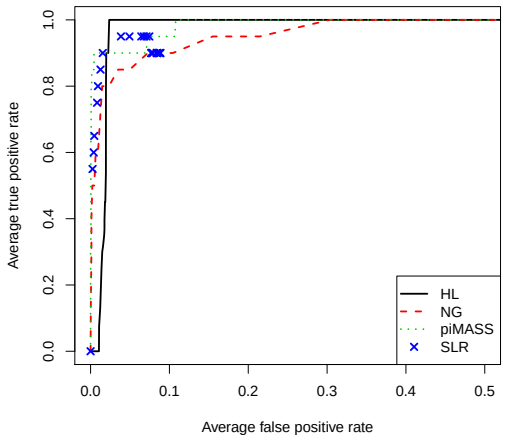
(a) Zooming ROC curve of the fifth scenario with 32000 cases and 32000 controls and $\frac{r^2}{r^2_{\max}} = 0.0005$.



(b) Zooming ROC curve of the sixth scenario with 32000 cases and 32000 controls and $\frac{r^2}{r^2_{\max}} = 0.01$.



(c) Zooming ROC curve of the seventh scenario with 32000 cases and 32000 controls and $\frac{r^2}{r^2_{\max}} = 0.71$.



(d) Zooming ROC curve of the eighth scenario with 32000 cases and 32000 controls and $\frac{r^2}{r^2_{\max}} = 0.32$.

Figure 4.14: ROC curves $(FPR \leq 0.5)$ for the stepwise logistic regression (SLR), PiMASS, Hyper lasso (HL) and the posterior credible interval using Normal-Gamma prior (NG) for an asymptotic normal likelihood. The methods are applied to 10 simulated dataset from Hapgen2 with two causal SNPs with odds ratios of 1.08 and 1.13 and having different MAFs (see Table 4.2). Each dataset has 32000 cases and 32000 controls. The NG MCMC was run for 20,000 iterations with 2,000 burn-in and thinning by 50. PiMASS was used with the default parameters. Hyper lasso was implemented with shape equal to 0.05 and scale equal to 0.004.

scenarios (see Figures 4.11(b) and 4.13(b)). Moreover, generally PiMASS seems the most constant approach in terms of FPR except in the third scenario it can be seen that the FPR for the datasets are varied and reached a high FPR. There is more variation in FPR at TPR = 1 for SLR and HL (often 0-1000) which is due to the hard decision nature of the approaches. Table 4.4 shows that the NG prior method seems to detect the causal SNPs at small FPRs compared to the HL. For example, in the fifth and sixth dataset of Scenario 1 the NG prior detects both causal SNPs at a very small FPR but the HL does not detect them in the most of datasets in Scenario 2.

Scenario 1								
Dataset	CI 1	CI 2	HL 1	HL 2	PiMASS 1	PiMASS 2	SLR 1	SLR 2
1	14	197	0	0	3	7	NA	3
2	14	197	0	0	3	7	NA	3
3	0	239	0	0	0	3	0	3
4	0	249	0	0	0	3	0	3
5	3	21	0	1000	0	3	0	0
6	3	21	0	1000	0	3	0	0
7	0	253	0	0	0	3	0	3
8	3	21	0	1000	0	3	0	0
9	0	256	0	0	0	3	0	3
10	0	69	0	0	0	0	3	24
Scenario 2								
Dataset	CI 1	CI 2	HL 1	HL 2	PiMASS 1	PiMASS 2	SLR 1	SLR 2
1	95	1000	NA	1000	0	248	15	1000
2	0	230	0	1000	0	4	4	1000
3	0	11	0	0	0	0	7	7
4	0	26	NA	1000	0	11	15	1000
5	0	29	NA	1000	0	11	15	1000
6	0	15	4	1000	4	7	15	15
7	0	328	NA	1000	0	4	11	1000
8	3	18	0	1000	0	4	0	11
9	4	595	NA	1000	0	7	4	1000
10	7	296	0	1000	7	11	11	1000

Table 4.4: $1000\times$ *FPR for Scenario 1 and Scenario 2 using the NG Credible Interval (CI), HL, PiMASS, and SLR all at TPR* $= 0.5$ *and TPR* $= 1$*. An NA indicates the FPR at the corresponding TPR is not available.*

Scenario 3								
Dataset	CI 1	CI 2	HL 1	HL 2	PiMASS 1	PiMASS 2	SLR 1	SLR 2
1	104	785	4	1000	0	960	0	1000
2	100	577	4	1000	0	960	0	1000
3	108	577	4	1000	0	960	0	1000
4	104	577	4	1000	0	960	0	1000
5	0	57	0	0	4	4	0	54
6	828	1000	NA	1000	14	65	29	1000
7	43	423	7	1000	11	65	14	1000
8	186	308	NA	1000	14	18	NA	1000
9	437	1000	NA	1000	36	122	NA	1000
10	262	344	4	4	11	25	4	7
Scenario 4								
Dataset	CI 1	CI 2	HL 1	HL 2	PiMASS 1	PiMASS 2	SLR 1	SLR 2
1	NA	4	NA	1000	4	7	NA	11
2	18	109	4	4	0	0	4	1000
3	4	7	0	0	0	0	42	1000
4	18	112	4	4	0	0	4	1000
5	18	116	4	4	0	0	4	1000
6	21	112	4	4	0	0	4	1000
7	7	453	4	1000	0	11	42	1000
8	7	446	4	1000	0	11	42	1000
9	7	453	4	1000	0	11	42	1000
10	NA	4	NA	1000	4	7	NA	11

Table 4.5: $1000\times$ *FPR for Scenario 3 and Scenario 4 using the NG Credible Interval (CI), HL, PiMASS, and SLR all at TPR* $= 0.5$ *and TPR* $= 1$*. An NA indicates the FPR at the corresponding TPR is not available.*

4.9 SNP selection by method

In this section we will discuss the performance of the pervious methods (NG, HL, SLR, and PiMASS) in terms of the selected SNPs. Are these methods selecting the same SNPs or does each method select different SNPs across the 10 datasets? We run and examine this in detail only for the first scenario and the fourth scenario because the level of LD for the fourth scenario is much stronger than the level of LD for the first scenario (see Table 4.1). The results are presented in Tables 4.8 and 4.9. For the SLR method we chose 1×10^{-5} as the univariate threshold rather than 5×10^{-8} which is the standard threshold employed by many fine mapping studies because the standard threshold selected no SNPs. For the HL method we set the shape to be $\lambda = 0.05$ and the scale to be $\gamma = 0.002$ (see Section 4.7). These values lead to selection of a small number of SNPs that have non-zero estimated effect sizes (19 SNPs were selected in total across all 10 datasets in Scenario 1). In the NG prior the size of credible interval chosen for all 10 datasets is 70%. This is because, although the 70% size is not able to detect the common causal SNP in Scenario 1, it leads to the selection of 19 SNPs in Scenario 1 and 26 SNPs in Scenario 4 in total across all 10 datasets which is comparable with the number of SNPs selected by HL and so makes for a fairer comparison of SNPs selected.

We will demonstrate and describe not only selection of SNPs for all methods (NG, HL, SLR, and PiMASS) in general, but also selection of SNPs within the LD block surrounding the common causal SNP (see Tables 4.6 and 4.7).

Tables 4.8 and 4.9 show the SNPs selected by the four methods (NG, HL, SLR, and PiMASS). Each table is divided into twelve columns: SNPs, r_1^2, r_2^2, MAF, NG, HL, SLR, PiMASS, NG-PiMASS, HL-PiMASS, and SLR-PiMASS. In the tables, SNPs represents the SNP ID, r_1^2 refers to the LD with the common causal SNP, r_2^2 refers to the LD with the rare causal SNP, MAF represents the minor allele frequency of each SNP selected, NG represents to the total number of times a SNP was selected out of the 10 datasets using a 70% credible interval, HL refers to the total number of times a SNP was selected out of the 10 datasets using HL, SLR shows the total number of times a SNP was selected out of the 10 dataset using SLR, PiMASS represents the mean of the PIP rank from PiMASS over the 10 datasets for the SNPs selected in at least one of NG, HL, or SLR, NG-PiMASS refers to the mean of the

SNP PIP rank in PiMASS including only the SNP PIP rank from datasets that NG selected, HL-PiMASS refers to the mean of the SNP PIP rank in PiMASS including only the SNP PIP rank from datasets that HL selected, and SLR-PiMASS represents the mean of the SNP PIP rank in PiMASS including only the SNP PIP rank from datasets that SLR selected. For example, SNP79 was selected twice with the 70% CI in the NG in Scenario 1 (see Table 4.8). The mean rank of the PIP from PiMASS across the 10 datasets for this SNP is approximately 229.8, whereas it is approximately 1 for only those two datasets (the first and second dataset) where the NG picked this SNP up.

The simulation parameters are from Scenario1 and Scenario 4 in Table 4.1. The blue SNP represents the common causal SNP and the red SNP represents the rare causal SNP. Within the PiMASS column the bold green colour represents the SNPs that are in the top 12 rank of PiMASS in at least one dataset. Note the other scenarios (Scenario 2, 3, 5, 6, 7, 8) were not considered.

4.9.1 Selection of SNPs in general

Here we will discuss the SNPs selected by the methods in general and will focus on selection of causal SNPs by each method.

Table 4.8 shows that in Scenario1 the NG approach was not able to detect the common causal SNP (SNP47) at the 70% size of CI in any of the 10 datasets. However, HL and SLR detected the common causal SNP in Scenario1 in 10 out of 10 datasets and 3 out of 10 datasets respectively. In addition, PiMASS selected the common causal SNP with a high rank of the PIP, with a mean rank over the 10 datasets of approximately 2.1, where 1 represents the top ranked SNP. Notice that SLR selected the common causal SNP in the fifth, sixth, and eighth dataset where its rank is the highest rank using PiMASS (see Table 4.10).

Table 4.9 shows that in Scenario 4 the common causal SNP (SNP46) was detected by the NG using a 70% CI in 2 out of 10 datasets. In this scenario HL selects SNP46 in 5 datasets whilst SLR never captures it. The mean PIP rank in PiMASS is 3.3.

In Scenario1 the rare causal SNP (SNP230) was detected in the NG using an 80% CI in 4 out of 10 datasets (results not shown). The datasets were the third, fourth, seventh, and ninth

datasets and the rare causal SNP had the highest PIP rank in these datasets showing similarity in SNP selection between NG and PiMASS (see Table 4.10). However, Table 4.8 shows that in Scenario1 relaxing the size of CI to 70% (which we are interested in) gave the NG more opportunity to detect the SNP from other datasets. With a 70% CI, the NG detected the rare causal SNP in 8 out of 10 (all the datasets except the first and the second datasets that both have the lowest PIP rank for SNP230 compared to the other 8 datasets). HL and SLR detected the rare causal SNP (SNP230) in 7 out of 10 datasets and 6 out of 10 datasets respectively in Scenario 1. In addition, in Scenario1 PiMASS selected the rare causal SNP (SNP230) with a mean rank across the 10 datasets of approximately 2.2 (i.e. approximately the second highest rank).

Table 4.9 shows that in Scenario 4 the rare causal SNP (SNP212) was detected in the NG with a 70% CI in 3 out of 10 datasets. SLR never captures the rare causal SNP in Scenario 4 whilst HL captures it in 8 datasets. The mean PIP rank in PiMASS is 1.2.

Table 4.8 shows that in Scenario1 HL and SLR both selected SNP198 in 2 datasets, although it has a low PIP rank. SNP156 was selected in 4 out of 10 datasets by SLR with the initial univariate threshold 1×10^{-5} and the mean PIP rank for those 4 datasets is 11 using PiMASS. However, Table 4.9 shows that in Scenario 4 HL was selecting many SNPs within the top 12 PIP rank from PiMASS in least one dataset (such as SNP17, SNP47, SNP84, SNP129, SNP143, and SNP146), but also one that was not (SNP134). The same is true about the NG in Scenario 1, many of the selected SNPs particularly at the 70% CI were within the top 12 rank PIP from PiMASS in least one dataset (SNP24, SNP73, SNP79, SNP129, SNP215, and SNP284) (see also Table 4.10). Additionally, in the fourth scenario 70% CI the NG selected SNP47, SNP84, and SNP146 that are all highly ranked by PiMASS (see Tables 4.9, and 4.11). PiMASS always places both the common casual SNP and the rare causal SNP within the top 4(5) PIP rank in Scenario1(Scenario4).

Table 4.11 shows that in the fourth scenario the PiMASS PIPs of the top ranked SNPs in some datasets are low. In 6 out of 10 datasets in Table 4.11 the maximum PIP is 0.12 or less and the maximum PIP of the top ranked SNPs is 0.62. However, in the first scenario in Table 4.10, the PiMASS PIPs of the top ranked SNPs in some datasets are high. In only 3 out of the 10 datasets in Table 4.10 is the maximum PIP 0.31 or less and the maximum PIP of the

top ranked SNPs is 0.84. So although PiMASS ranks the common and rare SNPs highly in all datasets, they often have very low PIP meaning these is relatively little confidence in these SNPs being in the model.

In Scenario 1 both the NG prior and the HL selected 19 SNPs each whereas the SLR selected only 15 SNPs. All methods are able to select common and rare SNPs. However, the NG prior selected rare SNPs more often than the others (see Table 4.8). In Scenario 4 again both the NG prior the HL selected approximately 30 SNPs whereas the SLR selected no SNPs. Both the NG prior and the HL are able to select common and rare SNPs. In both Scenarios 1 and 4, most of the SNPs selected by the NG prior, the HL and the SLR have high PIP rank. Also there are SNPs with high PIP rank not selected by the others such as SNP48 with a PIP rank of 3.

The criteria is to make both methods choose the same number of SNPs. This criteria might be not the best. Therefore, future research is to investigate a suitable criteria with which to compare these methods. Based on the criteria we considered it seems that the HL selected the potential causal SNPs better than the NG prior and the SLR in both Scenarios 1 and 4.

4.9.2 Selection of SNPs within the LD block

In section 4.9.1 we compared between the methods considered in term of the selection of SNPs in general using only Scenario 1 and 4. However, in this section we will concentrate on the selection of SNPs within the LD block around the common causal SNP. Also we will discuss the impact of the LD between the two potential causal SNPs on the selection of SNPs.

Table 4.9 illustrates the effect of LD on the performance of the HL in Scenario 4 which only detects the common causal SNP in 5 out of 10 datasets, whereas the HL detected the common causal SNP in 10 out of the 10 datasets in Scenario 1 (see Table 4.8). Moreover, in Scenario 4 (Table 4.9) , it can be seen that the common causal SNP (SNP46) is selected in 5 of the 10 datasets (from second dataset to sixth dataset) whereas SNP47 which is located in the LD block with an r^2 of 0.9 was selected in 3 out of the 10 datasets (from seventh dataset to ninth dataset). There is no intersection of these 3 datasets with the 5 datasets that selected the common causal SNP showing that the HL is selecting the common causal SNP or the SNP

most correlated with it but never both.

This impact of high LD extends also to SLR. Table 4.9 shows that SLR was not able to detect the common causal SNP (SNP46) at the initial univariate threshold of 1×10^{-5} in any datasets. However, Table 4.8 indicates that it was able to detect the common causal SNP (SNP47) in 3 out of the 10 datasets at the the same initial univariate threshold of 1×10^{-5}.

In Scenario 1 the NG prior did not select the common causal SNP (SNP47) or any SNPs located within the LD with it at 70% credible interval size. However, in Scenario 4 it did select the common causal SNP, SNP46, and the highest correlated SNP with it, SNP47, (see Tables 4.8 and 4.9). However, PiMASS gave high ranks to most of the SNPs within the LD mostly with a top 12 PIP rank in both Scenarios (see Tables 4.10 and 4.11).

It seems that the HL deals better than the other methods with selection of SNPs within the LD block in Scenario 1. In PiMASS, in both Scenarios the SNPs within the LD block obtained high PIP rank which means PiMASS distributed the signal between the SNPs within the LD block.

The 70% credible interval in the NG prior might be not a good choose for the NG prior to address the selection of SNPs within the LD block around the common causal SNP. Therefore, we will compare the SNPs in the LD block using the mean maximum credible interval size in both Scenarios. Moreover, we will give mean PIP for those SNPs.

Table 4.6 indicates the SNPs within the LD block $(r^2 \geq 0.8)$ in the first scenario. It can be seen that in several datasets PiMASS gave several SNPs within the LD block very similar PIP. For instance, Table 4.10 shows that in Scenario 1 the PIP for the causal common SNP (SNP47) is 0.35 in 4 datasets. The PIP for SNP49 is 0.34 in these 4 datasets. Therefore, PiMASS cannot distinguish them. Moreover, Table 4.6 shows that the mean PIP over the 10 datasets is 0.32 and 0.25 for the common causal SNP (SNP47) and SNP48 which have r^2 of 0.93 and in the first two datasets the PIP for SNPs 47 and 48 are 0.33 and 0.3 (see Table 4.10). Again PiMASS cannot distinguish them. However, Table 4.7 displays the SNPs within the LD block in the fourth scenario. It can be noticed that the mean PIP over the 10 datasets is 0.19 and 0.21 for the common causal SNP (SNP46) and SNP47 which have r^2 of 0.9. It can be seen that the mean PIP of the SNPs within the LD block in Scenario 1 is greater than the mean PIP in Scenario 4.

Although the NG prior with 70% credible interval never selected the common causal SNP in Scenario 1, the mean maximum credible interval size of the common causal SNP is the highest (42%) among the SNPs in LD with it (see Table 4.6). Moreover, Table 4.9 shows that in Scenario 4 the NG prior with 70% selected the common causal SNP (SNP46) in 2 out of the 10 datasets (first dataset, and tenth dataset) and selected SNP47 (which is located in the LD block at $r^2 = 0.9$) in 3 out of the 10 datasets (from seventh dataset to ninth dataset) again in mutually exclusive datasets. However, the mean maximum credible interval size of the common causal SNP is the highest (39.5%) among those in the LD block (see Table 4.7). It can be seen that in Scenario 4 the mean maximum credible interval size of the common causal SNP and the highest SNPs correlated with the common SNP are quite similar (39.5% and 37% respectively). Therefore, if there is LD between the two causal SNPs it looks likely that it may affect the selection of SNPs with the LD block.

SNPs	26	31	35	**47**	48	49	50	59	61	63	69	80
r^2	0.84	0.83	0.84	**1**	0.93	0.81	0.86	0.8	0.81	0.81	0.81	0.8
Mean maximum CI size	4.5%	4%	4.5%	42%	14%	24%	8%	9%	6.5%	11%	10%	5%
Mean PIP	0.18	0.17	0.17	**0.32**	0.25	0.20	0.19	0.16	0.17	0.17	0.17	0.18
Mean rank of PIP	21.6	32.7	26.5	**2.1**	10.8	22	20.95	52.6	38.1	23.9	41.8	27

Table 4.6: r^2 with the common causal SNP for SNPs in its LD block for Scenario 1 in Table 4.1, where the LD block is defined as $r^2 \geq 0.8$. SNP47 in blue is the common causal SNP.

4.10 Discussion

There are many statistical approaches to variables selection in regression models. In fine-mapping, we are also interested in selecting a model that includes the causal SNPs. The challenge within fine mapping is the high correlation between SNPs, which can lead to selection of a non-casual SNP in high LD with the casual SNP. In SLR, the LD affected the performance in that it was not able to select any SNPs at the initial univariate threshold 1×10^{-5}.

SNPs	25	30	34	**46**	47	48	58	65	71
r^2	0.83	0.81	0.83	**1**	0.9	0.84	0.8	0.82	0.82
Mean maximum CI size	4%	11.5%	14%	39.5%	37%	17%	10%	11%	15.5%
Mean PIP	0.11	0.11	0.11	**0.19**	0.21	0.11	0.10	0.10	0.10
Mean rank of PIP	22.2	15.6	20.1	**3.3**	21.7	28.2	61.3	110.3	96.15

Table 4.7: r^2 with the common causal SNP for SNPs in its LD block for Scenario 4 in Table 4.1, where the LD block is defined as $r^2 \geq 0.8$. SNP46 in blue is the common causal SNP.

The PiMASS approach gives a posterior inclusion probability (PIP) for each SNP. We have seen LD can lead to correlated SNPs having higher PIP than the causal SNP as we have seen in the fourth scenario (see Table 4.11). The HL, in the first scenario, seems to perform well in terms of selecting the causal SNPs (it selected the common SNP in 10 out of the 10 datasets and the rare causal SNP in 7 out of the 10 datasets). However, this is not the case in the fourth scenario where the HL selected some of the non-causal SNPs located within the LD block. The 70% posterior credible interval in the NG analysis did successfully select the rare causal SNP in most of datasets with less success for the common causal SNP. In Scenario 1, the HL and SLR invariably only select the causal SNP from the LD block, but the NG prior with 70% credible interval size did not select the common causal SNP at all. In this setting it seems the HL work reasonably well in terms of within-block localisation. However, in Scenario 4 the NG and the HL sometimes pick up the wrong SNP in the Scenario 4 whilst SLR selects none. PiMASS usually gave high rank to the causal SNPs but also it often selected SNPs within the LD block with a top 12 rank.

In both Scenarios 1 and 4, the mean maximum credible interval size of the common causal SNP is higher than any other SNPs within the LD block, although the common causal SNP was not selected in Scenario 1 at credible interval of size 70% and was selected less often than the highest SNP correlated with in Scenario 4. Therefore, setting the credible interval size for the NG prior to be a credible interval that select similar total number of HL select might be not a perfect choice.

SNPs	r_1^2	r_2^2	MAF	NG	HL	SLR	PiMASS	NG-PiMASS	HL-PiMASS	SLR-PiMASS
Scenario 1 (70% CI)										
SNP47	1.00	0.01	0.28	0	10	3	**2.10**	NA	2.10	1
SNP79	0.15	0.07	0.39	2	0	0	**229.8**	1	NA	NA
SNP117	0.01	0	0.01	2	0	0	97.50	13	NA	NA
SNP156	0.04	0.27	0.16	0	0	4	**134.1**	NA	NA	11
SNP198	0.01	0.36	0.2	0	2	2	101.9	NA	22	22
SNP215	0	0.03	0.02	4	0	0	**112.2**	2	NA	NA
SNP230	0.01	1	0.09	8	7	6	**2.20**	1.75	1.86	2
SNP284	0	0	0	3	0	0	**85**	9	NA	NA

Table 4.8: Total number of times a SNP was selected out of the 10 datasets in Scenario 1 using 70% credible intervals in the NG, HL and SLR. In addition, the mean PIP rank from PiMASS over all 10 datasets, and the mean PIP rank from PiMASS only over datasets selected by NG, HL, or SLR are reported. For more detail of table content (see Section 4.9).

One can ask a question about the utility of the rank of the PIP or the mean or median posterior effect size in the NG prior. The disadvantage is in selecting the proper threshold to apply which invariably relies on a rule of thumb. There is clearly an opportunity to use functional genomic information to improve localisation of the signal within the LD block in particular and to reduce the number of false positives. We will consider this aspect in Chapters 6-8.

One can be noticed that the priors for HL, PiMASS, and the NG prior might be resemble because all of them have mass close to zero and heavy tails elsewhere. The interesting question is that if all three methods have similar priors does this explain the similarities in the models or the results? We believe that it is quite difficult to obtain exactly same priors come from three different models, but they might be similar priors. Moreover, obtaining similar results between some methods may not mean they have similar prior because each method has its own procedure for selecting SNPs to be potential causal SNPs.

SNPs	r_1^2	r_2^2	MAF	NG	HL	SLR	PiMASS	NG-PiMASS	HL-PiMASS	SLR-PiMASS
Scenario 4 (70% CI)										
SNP17	0.18	0.01	0.07	0	2	0	**97.3**	NA	6	NA
SNP44	0.01	0	0.01	4	0	0	110	NA	NA	NA
SNP46	1	0.09	0.31	2	5	0	**3.3**	4	2	NA
SNP47	0.9	0.08	0.33	3	3	0	**21.7**	2	2	NA
SNP84	0	0	0	3	3	0	**42.1**	8	8	NA
SNP89	0.02	0.01	0.05	1	0	0	225.1	46	NA	NA
SNP129	0.03	0	0.01	0	4	0	**39.5**	NA	3	NA
SNP134	0.05	0.02	0.16	0	3	0	165.75	NA	15	NA
SNP143	0.01	0.18	0.37	0	2	0	**71.35**	NA	1	NA
SNP144	0	0	0	4	0	0	58.75	10	NA	NA
SNP146	0	0	0	2	2	0	**113.65**	NA	7	NA
SNP212	0.09	1	0.10	3	8	0	**1.2**	2	1	NA
SNP227	0.01	0.03	0.21	4	0	0	169.8	193	NA	NA

Table 4.9: Total number of times a SNP was selected out of the 10 datasets in Scenario 4 using 70% credible intervals in the NG, HL and SLR. In addition, the mean PIP rank from PiMASS over all 10 datasets, and the mean PIP rank from PiMASS only over datasets selected by NG, HL, or SLR are reported. For more detail of table content (see Section 4.9).

Scenario 1										
Dataset	1	2	3	4	5	6	7	8	9	10
Rank1	79	79	230	230	**47**	**47**	230	**47**	230	230
PIP	0.40	0.40	0.84	0.84	**0.31**	**0.31**	0.84	**0.31**	0.84	0.62
Rank2	**47**	**47**	215	215	*48*	*48*	215	*48*	215	**47**
PIP	**0.33**	**0.33**	0.40	0.40	*0.23*	*0.23*	0.40	*0.23*	0.40	**0.25**
Rank3	*48*	*48*	**47**	**47**	230	230	**47**	230	**47**	94
PIP	*0.30*	*0.30*	**0.35**	**0.35**	0.16	0.16	**0.35**	0.16	**0.35**	0.25
Rank4	230	230	*49*	*49*	24	24	*49*	24	*49*	68
PIP	0.25	0.25	*0.34*	*0.34*	0.14	0.14	*0.34*	0.14	*0.34*	0.24
Rank5	125	125	154	154	216	216	154	216	154	187
PIP	0.23	0.23	0.32	0.32	0.14	0.14	0.32	0.14	0.32	0.20
Rank6	94	94	*50*	*50*	129	129	*50*	129	*50*	154
PIP	0.21	0.21	*0.31*	*0.31*	0.11	0.11	*0.31*	0.11	*0.31*	0.19
Rank7	78	78	187	187	80	80	187	80	187	14
PIP	0.19	0.19	0.30	0.30	0.10	0.10	0.30	0.10	0.30	0.18
Rank8	73	73	289	289	*50*	*50*	289	*50*	289	281
PIP	0.19	0.19	0.30	0.30	*0.10*	*0.10*	0.30	*0.10*	0.30	0.18
Rank9	271	271	290	290	284	284	290	284	290	143
PIP	0.18	0.18	0.30	0.30	0.10	0.10	0.30	0.10	0.30	0.18
Rank10	244	244	114	114	*49*	*49*	114	*49*	114	22
PIP	0.17	0.17	0.29	0.29	*0.08*	*0.08*	0.29	*0.08*	0.29	0.17
Rank11	68	68	156	156	*61*	*61*	156	*61*	156	196
PIP	0.17	0.17	0.28	0.28	*0.08*	*0.08*	0.28	*0.08*	0.28	0.17
Rank12	21	21	35	35	*69*	*69*	35	*69*	35	83
PIP	0.17	0.17	0.28	0.28	*0.08*	*0.08*	0.28	*0.08*	0.28	0.17

Table 4.10: Top 12 ranked SNPs from PiMASS and the PIP for these SNPs for the first scenario in Table 4.1. The bold blue colour refers to the common causal SNP, the red colour represents the rare casual SNP, and the italic green colour refers to the SNPs within the LD block given in Table 4.6.

Scenario 4										
Dataset	1	2	3	4	5	6	7	8	9	10
Rank1	143	212	212	212	212	212	212	212	212	143
PIP	0.11	0.11	0.19	0.12	0.12	0.12	0.63	0.63	0.63	0.11
Rank2	212	**46**	**46**	**46**	**46**	**46**	*47*	*47*	*47*	212
PIP	0.11	**0.09**	**0.14**	**0.08**	**0.08**	**0.08**	*0.62*	*0.62*	*0.62*	0.11
Rank3	114	129	*47*	129	129	129	18	18	18	114
PIP	0.11	0.02	*0.09*	0.03	0.03	0.03	0.46	0.46	0.46	0.11
Rank4	**46**	*25*	52	*25*	*25*	*25*	35	35	35	**46**
PIP	**0.09**	*0.02*	0.05	*0.03*	*0.03*	*0.03*	0.45	0.45	0.45	**0.09**
Rank5	100	*34*	35	*34*	*34*	*34*	**46**	**46**	**46**	100
PIP	0.07	*0.02*	0.04	*0.03*	*0.03*	*0.03*	**0.42**	**0.42**	**0.42**	0.07
Rank6	17	*47*	22	*47*	*47*	*47*	52	52	52	17
PIP	0.07	*0.01*	0.03	*0.03*	*0.03*	*0.03*	0.37	0.37	0.37	0.07
Rank7	146	26	*71*	26	26	26	3	3	3	146
PIP	0.06	0.02	*0.03*	0.03	0.03	0.03	0.36	0.36	0.36	0.06
Rank8	262	107	110	107	107	107	84	84	84	262
PIP	0.06	0.02	0.03	0.02	0.02	0.02	0.36	0.36	0.36	0.06
Rank9	45	*48*	10	*48*	*48*	*48*	111	111	111	45
PIP	0.05	*0.01*	0.02	*0.02*	*0.02*	*0.02*	0.36	0.36	0.36	0.05
Rank10	110	144	114	144	144	144	60	60	60	110
PIP	0.05	0.01	0.02	0.02	0.02	0.02	0.36	0.36	0.36	0.05
Rank11	194	*30*	25	*30*	*30*	*30*	28	28	28	194
PIP	0.04	*0.03*	0.02	*0.02*	*0.02*	*0.02*	0.34	0.34	0.34	0.04
Rank12	109	83	223	83	83	83	143	143	143	109
PIP	0.04	0.01	0.02	0.02	0.02	0.02	0.34	0.34	0.34	0.04

Table 4.11: Top 12 ranked SNPs from PiMASS and the PIP for these SNPs for the fourth scenario in Table 4.1. The bold blue colour refers to the common causal SNP, the red colour represents the rare casual SNP, and the italic green colour refers to the SNPs within the LD block given in Table 4.7.

Chapter 5

Applying the four chosen methods to the iCOGs data

In Chapter 4 we compared the effectiveness of several methods in fine mapping our simulated data. Moreover, we discussed the variability between datasets for each method. Applying these methods on the simulation data allows us to compare the selection of the true causal SNPs but applying these methods on a real dataset is also informative and allows us to assess the between-method variation in SNPs selected. Therefore, the aim of this Chapter is to apply the methods on the iCOGs data (see section 5.1). In this Chapter we will describe the iCOGs data and compare and contrast the methods based on their performance.

5.1 iCOGs data

Recently, many studies were conducted using the iCOGs array that was designed by the Collaborative Oncological Gene-environment Study (COGS) for fine-mapping studies. These studies include SNPs that are highly associated with breast, ovarian and prostate cancer. The iCOGs array has 211,155 SNPs (Michailidou et al., 2013b). However, we are interested in a particular gene implicated in breast cancer. Therefore, we concentrate only on the region located on Chromosome 2 between base positions 201500074 and 202569992 that includes the CASP8 gene. There were already 585 SNPs genotyped in breast cancer case-control samples from the Breast Cancer Association Consortium. We will apply the methods on a total

1733 SNPs with 46450 cases and 42500 controls (total sample size is 89050). Those 1733 SNPs come from two sources: 501 SNPs are selected out of the 585 SNPs from the Breast Cancer Association Consortium that passed quality control checks, the other 1232 SNPs were selected using IMPUTE2 (Marchini and Howie, 2010). The SNPs were considered to be imputed successfully if their imputation accuracy was greater than 90%.

5.1.1 Preparing the iCOGs data

We extracted the 1733 SNPs from the raw data. The format of the data was not suitable for the methods we use because the data was formatted such that each individual has a value for each possible genotype. For example, if A and B are two alleles, then the possible genotype are: AA, AB, and BB. The raw data of the 501 genotyped SNPs give a binary value (e.g 0, 0, and 1) for AA, AB, and BB counts respectively. We need to code this data to be 2, 1, and 0 based on the number of copies of the minor allele (which is A in this case). In addition, the raw data of the 1232 SNPs that were imputed contained a rational number (e.g 0.1, 0.2, and 0.7) for AA, AB, and BB counts respectively. This reflects the uncertainty in the imputation process. We need to code this based on the number of copies of the minor allele. To achieve this, we calculated the expected value of the number of copies of the minor allele A given by

$$p_{(AA)} \times 2 + p_{(AB)} \times 1 + p_{(BB)} \times 0, \tag{5.1}$$

where $p_{(AA)}$ is the probability of the genotype AA, $p_{(AB)}$ is the probability of the genotype AB, and $p_{(BB)}$ is the probability of the genotype BB. Table 5.1 shows examples of transforming the raw data from 3 genotype values to the expected value.

SNPs	(2) AA	(1) AB	(0) BB	$E\,(A)$
1	1	0	0	2
2	0	0	1	0
3	0	1	0	1
4	0.1	0.2	0.7	0.4

Table 5.1: Calculating the expected number of copies of the minor allele for an individual at 4 SNPs where the first three are typed SNPs and the last SNP is an imputed SNP.

5.2 Comparison tools

In simulation data scenarios, the potential causal SNPs can be specified perviously. Therefore, the performance of methods can be compared by ROC curves. However, in the real data this is not the case because the potential causal SNPs are not known. The aim here is to compare between the four methods: SLR, HL, piMASSS, and the NG prior in terms of the selected SNPs.

Here, the SLR was applied using the initial univariate threshold 1×10^{-5} rather than 5×10^{-8} which is the standard threshold employed by many fine-mapping studies because the standard threshold selected no SNPs. The HL method was applied with shape 0.05, because it is the smallest value that can be chosen in the HL , and scale 0.001 which corresponds to a FWER of 0.05. PiMASS was applied with the default values of all parameters to calculate a posterior inclusion probabilities. In PiMASS there is not a common rank threshold to use. Therefore, here we will consider that a SNP with PIP rank less than the total number of SNPs selected by the others is considered as a potential causal SNPs in PiMASS. A non-integer rank means a tie in the PIP. This is most likely to occur when two SNPs are in strong LD. The NG prior was applied to the iCOGs data using MCMC with 20000 iterations a burn-in of 2000 and thinned by 20 so the final thinned chain is 1000 observations long. The MCMC run for the iCOGs data took a week in Iceberg (the University computer cluster). The diagnostic convergence tests (trace plot, ACF plot and $\hat{R}$) were applied and all the parameter distributions converged (see Figures A.1 and A.2 in Appendix A). The credible interval was used as a posterior summary statistic. For the NG prior, the credible interval will be considered at four percentages: $85\%, 90\%, 95\%,$ and 99% because we are interested in comparison between different credible interval sizes.

5.3 Results and discussion

In this section we will discuss the results that are obtained from applying the methods (SLR, HL, piMASSS, and the NG prior) to the iCOGs data (see section 5.1). Moreover, we will discuss the characteristics of the SNPs that are selected by the NG prior but not the others.

The important characteristics that can be used to compare between the SNPs are the minor allele frequency and the r^2. However, the iCOGs data is available with only the minor allele frequency and calculating r^2 for the iCOGs data needs a lot of time. Therefore, we will compare between the selected SNPs based on their MAF.

Table 5.2 shows the SNPs selected by either SLR, HL or an 85% credible interval of the posterior effect sizes obtained from the NG. Moreover, the ranking of the posterior inclusion probabilities (PIP) from PiMASS were given for each selected SNP from one of the 3 methods. It can be seen that 18 SNPs were selected using the NG prior that were not selected by SLR or HL. However, the majority of the selected SNPs using the NG prior have top 50 PIP rank (except SNP6, SNP9, SNP122, SNP795, and SNP1608). It seems that the NG prior selects similar SNPs to PiMASS. However, the SLR and HL agree on only SNP1, and the PIP rank of SNP1 is 40 which is within the highest 3% of ranks. SNP31 is selected only with the HL method but the PIP rank of this SNP is 10.

Tables 5.3, 5.4, and 5.5 present the same results as Table 5.2 but with a 90%, 95%, and 99% NG credible interval respectively. As expected the number of selected SNPs with the NG prior decreases as the credible interval increases (13 selected SNPs with 90% CI, 8 selected SNPs with 95% CI and only 5 selected SNPs with 95% CI). It can be seen that the SNPs selected by the NG with 99% CI all have high PIP rank.

Table 5.2 shows that the NG prior selects SNPs that no other methods have selected: SNP122, SNP342, SNP795, SNP1177, SNP1241, SNP1274 and SNP1608. These SNPs include both common and rare SNPs. Moreover, it can be seen that the NG prior has ability to select the rare SNPs more than HL and SLR. For example it selects SNP122, SNP740 and SNP765 with MAF $0.05, 0.04$ and 0.09. SNP122 was selected only by the NG prior and has less PIP rank. The majority of SNPs selected by both the NG prior and PiMASS are common SNPs (SNP589, SNP740, SNP765, SNP990, SNP1056, SNP1101, SNP1146, SNP1159, SNP1244, SNP1245, SNP1253). It seems the SLR has difficulty with selecting rare SNPs because all SNPs selected by SLR are common SNP1, SNP2 SNP6, and SNP9 with MAF $0.13, 0.13, 0.26, 0.26$ respectively. Although there are only two SNPs were selected by the HL, they are both common and rare SNPs. Therefore, the HL has the ability to select both common and rare SNPs.

In conclusion each method has its own procedure for selecting SNPs to be potential causal SNPs. However, it may be more logical if one combines results between the methods so that the SNPs that are selected by several methods would be most likely to be potential causal SNPs than SNPs that are selected by only one method for example. Therefore, in the iCOGs data if one is interested in selecting potential causal SNPs, those SNPs which are selected by more than one method are candidates such as SNP31 and SNP589.

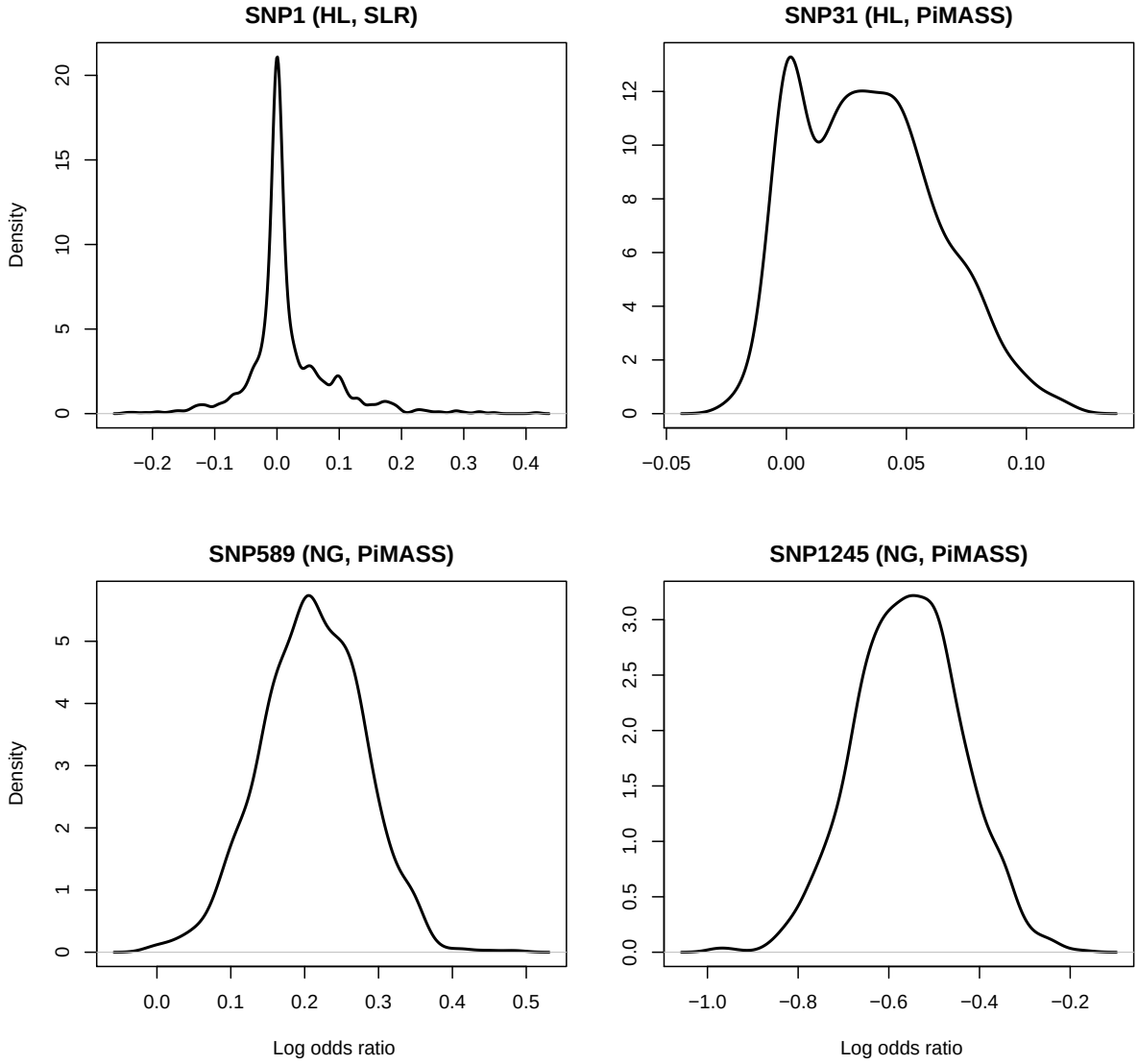


Figure 5.1: Density plots of posterior effect sizes from the Normal-Gamma prior for four selected SNPs by either SLR, HL, or a 99% *CI of the NG. When applied to the iCOGs data with 46450 cases and 42500 controls with 1733 SNPs. The individual plot title indicates which method selected the SNP. For PiMASS this equates to the SNP being in the top 13 PIP ranks.*

Figure 5.1 shows the posterior densities of the effect sizes for four typical selected SNPs with a NG 99% CI (from Table 5.5). SNP1 is selected by both HL and SLR but not the NG and has only moderate PIP rank (rank 40). SNP31 is selected by HL only but has a high PIP rank (rank 10). SNPs 589 and 1245 are selected only by the NG but also have very high PIP ranks (rank 7 and 2.5 respectively). It can be seen that the posterior densities of the two selected SNPs using the NG (SNP589 and SNP1245) have very little mass around zero. For SNP31 (chosen by HL and PiMASS) there is a lot of mass in the positive tail but there is also a lot of mass around zero. However, for SNP1 that is selected by SLR and HL only, most of the mass is very close to zero.

SNPs	MAF	NG	HL	SLR	PiMASS
SNP1	0.13	0	1	1	40
SNP2	0.13	0	0	1	32
SNP6	0.26	0	0	1	1411
SNP9	0.26	0	0	1	1351
SNP31	0.08	0	1	0	**10**
SNP122	0.05	1	0	0	606
SNP342	0.46	1	0	0	43
SNP589	0.20	1	0	0	**7**
SNP740	0.04	1	0	0	**23**
SNP765	0.09	1	0	0	**2.5**
SNP795	0.45	1	0	0	1376.5
SNP990	0.44	1	0	0	**8**
SNP1056	0.43	1	0	0	**19**
SNP1101	0.33	1	0	0	**2.5**
SNP1146	0.33	1	0	0	**22**
SNP1159	0.33	1	0	0	**5**
SNP1177	0.32	1	0	0	46
SNP1241	0.41	1	0	0	45
SNP1244	0.17	1	0	0	**2.5**
SNP1245	0.17	1	0	0	**2.5**
SNP1253	0.41	1	0	0	**18**
SNP1274	0.20	1	0	0	41
SNP1608	0.49	1	0	0	1671

Table 5.2: SNPs selected using an 85% credible interval of the posterior effect sizes in the NG prior or by HL or SLR along with the rank of the PiMASS PIP of the selected SNPs. We used 1 and 0 for NG, HL and SLR to indicate whether the SNP in a particular method was selected or not respectively. The bold green SNPs in the PiMASS column represents those SNPs in top 23 by PIP rank. It is applied to the iCOGs data with 1733 SNPs and a total sample size of 89050.

SNPs	MAF	NG	HL	SLR	PiMASS
SNP1	0.13	0	1	1	40
SNP2	0.13	0	0	1	32
SNP6	0.26	0	0	1	1411
SNP9	0.26	0	0	1	1351
SNP31	0.08	0	1	0	**10**
SNP122	0.05	1	0	0	606
SNP342	0.46	1	0	0	43
SNP589	0.20	1	0	0	**7**
SNP765	0.09	1	0	0	**2.5**
SNP990	0.44	1	0	0	**8**
SNP1101	0.33	1	0	0	**2.5**
SNP1159	0.33	1	0	0	**5**
SNP1177	0.32	1	0	0	46
SNP1241	0.41	1	0	0	45
SNP1244	0.17	1	0	0	**2.5**
SNP1245	0.17	1	0	0	**2.5**
SNP1253	0.41	1	0	0	**18**
SNP1274	0.20	1	0	0	41
SNP1608	0.49	1	0	0	1671

Table 5.3: SNPs selected using an 90% credible interval of the posterior effect sizes in the NG prior or by HL or SLR along with the rank of the PiMASS PIP of the selected SNPs. We used 1 and 0 for NG, HL and SLR to indicate whether the SNP in a particular method was selected or not respectively. The bold green SNPs in the PiMASS column represents those SNPs in top 19 by PIP rank. It is applied to the iCOGs data with 1733 SNPs and a total sample size of 89050.

SNPs	MAF	NG	HL	SLR	PiMASS
SNP1	0.13	0	1	1	40
SNP2	0.13	0	0	1	32
SNP6	0.26	0	0	1	1411
SNP9	0.26	0	0	1	1351
SNP31	0.08	0	1	0	**10**
SNP122	0.05	1	0	0	606
SNP342	0.46	1	0	0	43
SNP589	0.20	1	0	0	**7**
SNP765	0.09	1	0	0	**2.5**
SNP990	0.44	1	0	0	**8**
SNP1101	0.33	1	0	0	**2.5**
SNP1244	0.17	1	0	0	**2.5**
SNP1245	0.17	1	0	0	**2.5**

Table 5.4: SNPs selected using an 95% *credible interval of the posterior effect sizes in the NG prior or by HL or SLR along with the rank of the PiMASS PIP of the selected SNPs. We used 1 and 0 for NG, HL and SLR to indicate whether the SNP in a particular method was selected or not respectively. The bold green SNPs in the PiMASS column represents those SNPs in top 13 by PIP rank. It is applied to the iCOGs data with 1733 SNPs and a total sample size of 89050.*

SNPs	MAF	NG	HL	SLR	piMASS
SNP1	0.13	0	1	1	40
SNP2	0.13	0	0	1	32
SNP6	0.26	0	0	1	1411
SNP9	0.26	0	0	1	1351
SNP31	0.08	0	1	0	**10**
SNP589	0.20	1	0	0	**7**
SNP765	0.09	1	0	0	**2.5**
SNP1101	0.33	1	0	0	**2.5**
SNP1244	0.17	1	0	0	**2.5**
SNP1245	0.17	1	0	0	**2.5**

Table 5.5: SNPs selected using an 99% *credible interval of the posterior effect sizes in the NG prior or by HL or SLR along with the rank of the PiMASS PIP of the selected SNPs. We used 1 and 0 for NG, HL and SLR to indicate whether the SNP in a particular method was selected or not respectively. The bold green SNPs in the PiMASS column represents those SNPs in top 10 by PIP rank. It is applied to the iCOGs data with 1733 SNPs and a total sample size of 89050.*

Chapter 6

Incorporating the FS score into the normal gamma prior

In the previous chapters we discussed multivariate statistical methods that attempt to identify causal SNPs (Step wise logistic regression, Hyper lasso, PiMASS, and Normal-Gamma prior) in terms of their performance in fine-mapping studies with case-control data using ROC curves. We discussed in detail the NG prior as a continuous prior and its performance in selecting the causal SNP within an LD block.

As a result of the diversity and the abundance of functional genomic information, some researchers exploit this data by incorporating it into their models to prioritise causal SNPs in fine-mapping studies. Examples include p-values weighting (Saccone et al., 2008), a Bayesian Latent variable model (Fridley et al., 2011), Probabilistic Annotation Integrator (Kichaev et al., 2014), and Empirical Bayes Approach (Spencer et al., 2016) that uses Encode data to inform the prior probability of association in a Bayes factor approach. The aim of this chapter is to discuss how one can incorporate functional genomic information into the NG prior. We will focus in this project only on incorporating so-called functional significance (FS) scores into the NG prior.

6.1 Functional significance (FS) scores

There is an abundance of sources of functional genomic information including the F-SNP database (Lee and Shatkay, 2009) and the ENCODE database (ENCODE, 2011). Although functional genomic data has a limitation, in that it cannot provide information for all SNPs across the genome, it can be used in a Bayesian analysis in order to inform the prior distribution of effect sizes.

The F-SNP database gives functional significance (FS) scores that estimate the deleterious effects of SNPs (Lee and Shatkay, 2009). In order to calculate FS scores they applied several steps: at the beginning they retrieved predicted labels of the SNPs predicted from the public available dataset to be "deleterious" or "non-deleterious" and they classified the SNPs based upon four major genetic functional aspects: splicing, transcription, translation and post-translational modification. Then they calculated the tools reliability score by calculating the conditional probabilities of a deleterious SNP given that the tool predicts this SNP as deleterious. Each tool reports the confidence scores with different scales. Therefore, Lee and Shatkay (2009) normalise the confidence scores to take a value lying in the interval $[0, 1]$. They notice that the confidence score for a deleterious SNP takes a value between $[0.5, 1]$, whereas the confidence score for a non-deleterious SNP takes a value between $[0, 0.5]$. Moreover, they determined, for the SNPs in regulatory regions, whether they are conserved across multiple species or not, and if the SNPs are not within a conserved region the confidence score is set to be 0.5, because of the uncertainty in the functionality of the SNP. Ultimately, the FS scores are calculated based on the SNPs confidence scores (Lee and Shatkay, 2009). Note that there are also many SNPs with no FS scores that are considered as missing data, and are most likely to be non-deleterious (assuming deleterious and non-deleterious SNPs have missing information with the same per-SNP probability).

In this chapter, we will discuss incorporating the functional significance (FS) scores into the normal-gamma (NG) prior for the effect size. Here we propose to divide the FS scores into four groups based on the nature of scores rather than incorporate the acute value of FS scores which can be addressed in the future studies.

6.1.1 Incorporating the functional significance scores into the prior for the effect size

Bayesian approaches are very useful in fine mapping case-control studies because they allow the inclusion of information (the functional significance (FS) scores) into the prior. In our project, we use the normal-gamma prior and modify this prior to incorporate the published FS scores (Lee and Shatkay, 2009).

Recently, the functional significance (FS) scores were included into the NG prior (Boggis et al., 2016). However, these authors applied it to eQTL data (gene expression data) in which there was little LD between the SNPs. Also they partitioned the SNPs into seven groups based on their functional genomic information: Intergenic FS score, Intronic FS score, UTR3 FS score, Splicing FS score, Other FS score, Synonymous FS score, and Non-synonymous FS score. However, here we will incorporate the FS scores into the NG prior for use in fine mapping case-control studies with extremely high LD. Moreover, we will partition the SNPs according to their FS scores and not into pre-defined functional groups. This means we need to deal with SNPs that have no assigned FS score.

Using FS scores in the prior

Griffin and Brown (2010) proposed a prior for the hyper-parameters λ and $\gamma^{-2}|\lambda$ to provide variability around the Lasso prior ($\lambda = 1$). They used a single parameter for λ and γ^{-2} for all SNPs. We propose to modify the structure of the NG prior by allowing λ and γ^{-2} to take four different values based on four different classes of the confidence scores that the FS scores calculated from. Therefore, the FS scores were divided into four groups: FS > 0.5, FS $= 0.5$, FS < 0.5, and missing FS (FS = NA). We decided to classify the FS scores into four groups because Lee and Shatkay (2009) considered that SNPs with FS > 0.5 are most likely to be deleterious SNPs, SNPs with FS < 0.5 are unlikely to be deleterious SNPs, SNPs with FS $= 0.5$ are with a lack of evidence to be deleterious, and the deleterious effects of SNPs with FS = NA are unknown.

To do this, we propose to modify the distribution of the prior variance of beta for each group, where the effect size variance of SNPs in Group 1 has a distribution with the largest

expectation, the variance distribution of Group 3 SNPs has the lowest expectation, and Group 2 and Group 4 SNPs have a variance distribution that is a mixture of the Group 1 and Group 3 variance distributions. We set the variance distributions for each group as follows

$$\text{Group 1} \;:\; 2\lambda_1\gamma_1^2 \sim IG\left(2, M_1\right), \tag{6.1}$$

$$\text{Group 2} \;:\; 2\lambda_2\gamma_2^2 \mid w \sim w \times IG\left(2, M_1\right) + (1-w) \times IG\left(2, M_2\right), \tag{6.2}$$

$$\text{Group 3} \;:\; 2\lambda_3\gamma_3^2 \sim IG\left(2, M_2\right), \tag{6.3}$$

$$\text{Group 4} \;:\; 2\lambda_4\gamma_4^2 \mid h \sim h \times IG\left(2, M_1\right) + (1-h) \times IG\left(2, M_2\right), \tag{6.4}$$

where $M_1 = 0.01$, $M_2 = 0.001$ (see Section 6.2 for a justification of these values) and w and h are the mixture parameters for Group 2 and Group 4. In order to calculate the distributions of $\gamma_j^{-2} \mid \lambda_j, w, h$, a transformation was used and the distributions are given either gamma distributions or a mixture of gamma distributions.

For example, let us apply the relevant transformation for Equation 6.2.

$$f\left(2\lambda_2\gamma_2^2 \mid w\right) = w \times \frac{M_1^2}{\Gamma(2)}\left(2\lambda_2\gamma_2^2\right)^{-2-1} e^{-\frac{M_1}{2\lambda_2\gamma_2^2}} + (1-w) \times \frac{M_2^2}{\Gamma(2)}\left(2\lambda_2\gamma_2^2\right)^{-2-1} e^{-\frac{M_2}{2\lambda_2\gamma_2^2}} \tag{6.5}$$

$$\text{Let } \delta = \frac{1}{2\lambda_2\gamma_2^2} \Rightarrow 2\lambda_2\gamma_2^2 = \frac{1}{\delta} \text{ and } \mid \frac{d}{d\delta}\frac{1}{\delta} \mid = \frac{1}{\delta^2} \tag{6.6}$$

Standard techniques for transforming random variables and equation 6.5 gives

$$f\left(\delta \mid w\right) = w \times \frac{M_1^2}{\Gamma(2)}\left(\frac{1}{\delta}\right)^{-2-1}\left(\frac{1}{\delta}\right)^2 e^{-M_1\delta} + (1-w) \times \frac{M_2^2}{\Gamma(2)}\left(\frac{1}{\delta}\right)^{-2-1}\left(\frac{1}{\delta}\right)^2 e^{-M_2\delta}$$

$$\text{Thus, } f\left(\delta \mid w\right) = w \times \frac{M_1^2}{\Gamma(2)}\left(\delta\right)^3\left(\delta\right)^{-2} e^{-M_1\delta} + (1-w) \times \frac{M_2^2}{\Gamma(2)}\left(\delta\right)^3\left(\delta\right)^{-2} e^{-M_2\delta} \tag{6.7}$$

$$f\left(\delta \mid w\right) = w \times \frac{M_1^2}{\Gamma(2)}\left(\delta\right)^{2-1} e^{-M_1\delta} + (1-w) \times \frac{M_2^2}{\Gamma(2)}\left(\delta\right)^{2-1} e^{-M_2\delta} \tag{6.8}$$

$$f\left(\frac{1}{2\lambda_2\gamma_2^2} \mid w\right) = w \times \frac{M_1^2}{\Gamma(2)}\left(\frac{1}{2\lambda_2\gamma_2^2}\right)^{2-1} e^{-\frac{M_1}{2\lambda_2\gamma_2^2}} + (1-w) \times \frac{M_2^2}{\Gamma(2)}\left(\frac{1}{2\lambda_2\gamma_2^2}\right)^{2-1} e^{-\frac{M_2}{2\lambda_2\gamma_2^2}} \tag{6.9}$$

Therefore, it follows that

$$\frac{1}{2\lambda_2\gamma_2^2} \mid w \;\; \sim \;\; w \times Ga\left(2, M_1\right) + (1-w) \times Ga\left(2, M_2\right) \tag{6.10}$$

Properties of transforming a $Ga\left(.\right)$ distribution by a constant then give

$$\gamma_2^{-2} \mid \lambda_2, w \;\; \sim \;\; w \times Ga\left(2, \frac{M_1}{2\lambda_2}\right) + (1-w) \times Ga\left(2, \frac{M_2}{2\lambda_2}\right). \tag{6.11}$$

The same transformations can be applied to Equations 6.1, 6.3, and 6.4. The prior for the for $\gamma_j^{-2} \mid \lambda_j, w, h$ can be given as follows

$$\begin{aligned}
f(\gamma_1^{-2} \mid \lambda_1) &= \frac{(M_1/2\lambda_1)^2}{\Gamma(2)} \left(\gamma_1^{-2}\right)^{2-1} \exp\left(-\frac{M_1}{2\lambda_1}\gamma_1^{-2}\right), && (6.12)\\
f(\gamma_2^{-2} \mid \lambda_2, w) &= w \times \frac{(M_1/2\lambda_2)^2}{\Gamma(2)} \left(\gamma_2^{-2}\right)^{2-1} \exp\left(-\frac{M_1}{2\lambda_2}\gamma_2^{-2}\right) \\
&\quad + (1-w) \times \frac{(M_2/2\lambda_2)^2}{\Gamma(2)} \left(\gamma_2^{-2}\right)^{2-1} \exp\left(-\frac{M_2}{2\lambda_2}\gamma_2^{-2}\right) && (6.13)\\
f(\gamma_3^{-2} \mid \lambda_3) &= \frac{(M_2/2\lambda_3)^2}{\Gamma(2)} \left(\gamma_3^{-2}\right)^{2-1} \exp\left(-\frac{M_2}{2\lambda_3}\gamma_3^{-2}\right), && (6.14)\\
f(\gamma_4^{-2} \mid \lambda_4, h) &= h \times \frac{(M_1/2\lambda_4)^2}{\Gamma(2)} \left(\gamma_4^{-2}\right)^{2-1} \exp\left(-\frac{M_1}{2\lambda_4}\gamma_4^{-2}\right) \\
&\quad + (1-h) \times \frac{(M_2/2\lambda_4)^2}{\Gamma(2)} \left(\gamma_4^{-2}\right)^{2-1} \exp\left(-\frac{M_2}{2\lambda_4}\gamma_4^{-2}\right). && (6.15)
\end{aligned}$$

The prior distributions for w and h were selected to follow Beta distributions because it is conjugate with the Gamma distribution. Moreover, the parameters for the Beta distributions were selected based on the nature of the FS scores in the groups. The prior for w is a Beta distribution with shape 2 and scale 2, because w represents the weight parameter for Group 2 (FS $= 0.5$) in which SNPs are considered approximately equally likely to be deleterious or non-deleterious. In addition, the prior for h is a Beta distribution with shape 1 and scale 4, because h represents the weight parameter for Group 4 (FS $= NA$) in which SNPs are considered more likely not to be deleterious (see Figure 6.1) so that the mixture weight for the $IG\left(2, M_1\right)$ distribution should a *priori* be small.

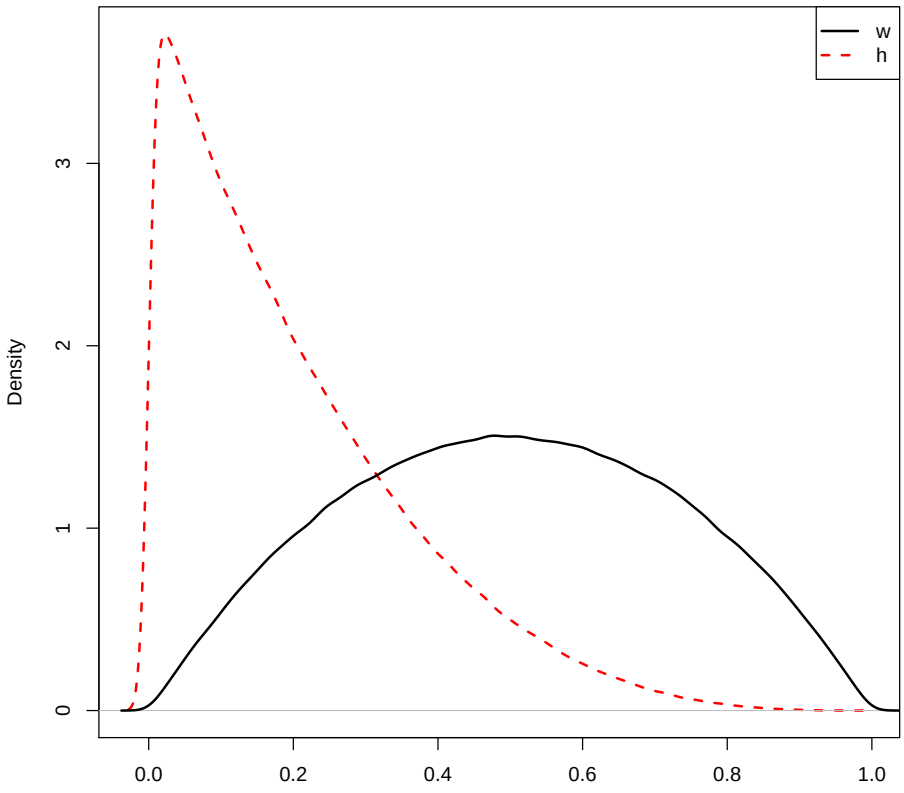


Figure 6.1: Prior densities for mixture components w and h.

The structure of the NG prior changes as follows. Let $j \in \{1, 2, 3, 4\}$ represent the FS score group and i represent the SNP label within the group with $i \in \{1, 2, \ldots, p_j\}$. Thus, β_{ij} is the effect size for the ith SNP in the jth group and ψ_{ij} is idiosyncratic prior variance.

$$f\left(\beta_{ij} \mid \psi_{ij}\right) = \frac{1}{\sqrt{2\pi|\psi_{ij}|}} \exp\left\{-\frac{1}{2}\frac{(\beta_{ij})^2}{\psi_{ij}}\right\} \tag{6.16}$$

$$f\left(\psi_{ij} \mid \lambda_j, \gamma_j^{-2}\right) = \frac{\left(\frac{1}{2\gamma_j^2}\right)^{\lambda_j}}{\Gamma(\lambda_j)} (\psi_{ij})^{\lambda_j - 1} \exp\left\{-\frac{\psi_{ij}}{2\gamma_j^2}\right\} \tag{6.17}$$

$$f\left(\gamma_j^{-2} \mid \lambda_j, w, h\right) \quad \text{see} \quad \text{Equations (6.12)} - \text{(6.15)} \tag{6.18}$$

$$f\left(\lambda_j\right) = 142.85 \times \exp\left(-142.85\lambda_j\right) \text{ see Section } 4.3 \tag{6.19}$$

$$\begin{aligned} w &\sim Beta\left(2,2\right), \\ f(w) &= \frac{\Gamma\left(2+2\right)}{\Gamma\left(2\right)\Gamma\left(2\right)} w^{2-1}\left(1-w\right)^{2-1}, \end{aligned} \tag{6.20}$$

$$\begin{aligned} h &\sim Beta\left(1,4\right), \\ f(h) &= \frac{\Gamma\left(1+4\right)}{\Gamma\left(1\right)\Gamma\left(4\right)} h^{1-1}\left(1-h\right)^{4-1}, \end{aligned} \tag{6.21}$$

where $M_1 = 0.01$ and $M_2 = 0.001$ (see Section 6.2).

6.2 Selecting M_1 and M_2

The basic idea that we propose to incorporate the FS scores into the NG prior, is to decrease the amount of shrinkage in the groups that have SNPs with high FS scores (Group 1 and Group 2) and increase the amount of shrinkage in the groups that has SNPs with low FS scores (Group 3 and Group 4). In order to achieve this one has to select a large variance for the NG prior variance (M_1) for the groups with high FS scores to allow more mass away from zero and select a smaller NG prior variance (M_2) for the groups with low FS scores to allow more mass close to zero. The aim is to select M_1 and M_2 to have reasonable amounts of shrinkage that lead to neither too much shrinkage nor too little shrinkage. We could of course shrink the posterior effect sizes in the SNP group with low FS scores very close to zero but if a causal SNP were placed in this group there would have be little chance of a convincing credible interval not containing zero.

To investigate the influence of the variance of the NG prior we consider the effect of different prior variances on the breast cancer top hits (BCTH) in a univariate analysis. We calculate the relative univariate shrinkage factor (SF) as follows

$$\text{SF} = 1 - \frac{\mathbb{E}_2\left(\beta \mid \hat{\beta}\right)}{\mathbb{E}_1\left(\beta \mid \hat{\beta}\right)}, \tag{6.22}$$

where $\mathbb{E}_1\left(\beta \mid \hat{\beta}\right)$ represents the mean of the posterior distribution with the NG prior with a variance of M_1, and $\mathbb{E}_2\left(\beta \mid \hat{\beta}\right)$ represents the mean of the posterior distribution with the NG prior with a variance of M_2. The SF can take any values either positive or negative. Note that SF close to 0 implies similar posterior means whilst values close to 1 imply much more shrinkage with the prior variance of M_2 compared to M_1.

The posterior mean $\mathbb{E}\left(\beta_i \mid \hat{\beta}_i\right)$ can be approximated as

$$\begin{aligned}
\mathbb{E}\left(\beta_i \mid \hat{\beta}_i\right) &= \int \beta_i f\left(\beta_i \mid \hat{\beta}_i\right) d\beta_i && (6.23)\\
&= \int \beta_i \frac{\pi\left(\beta_i\right) f\left(\hat{\beta}_i \mid \beta_i\right)}{f\left(\hat{\beta}_i\right)} d\beta_i && (6.24)\\
&= \frac{1}{f\left(\hat{\beta}_i\right)} \int \beta_i \pi\left(\beta_i\right) f\left(\hat{\beta}_i \mid \beta_i\right) d\beta_i && (6.25)\\
&\approx \frac{1}{n f\left(\hat{\beta}_i\right)} \sum_{k=1}^{n} \beta_k f\left(\hat{\beta}_i \mid \beta_k\right), && (6.26)
\end{aligned}$$

where β_k is sampled from the NG prior, n represents the number of the samples, $\pi\left(\beta_i\right)$ represents the prior of SNP β_i, $f\left(\hat{\beta}_i\right)$ represents the marginal likelihood, and $f\left(\hat{\beta}_i \mid \beta_i\right)$ represents the likelihood and is distributed asymptotically as a normal distribution with mean β_i and variance V. Note that Sequential Monte Carlo (SMC) was used to simulate the values of β_k from the NG prior in Equation 6.26. Note also that $f\left(\hat{\beta}_i\right)$ can be approximated as $\frac{\sum f\left(\hat{\beta}|\beta\right)}{n}$ and thus

$$\mathbb{E}\left(\beta_i \mid \hat{\beta}_i\right) = \frac{\sum_k \beta_k f\left(\hat{\beta}_i \mid \beta_k\right)}{\sum_k f\left(\hat{\beta}_i \mid \beta_k\right)}. \quad (6.27)$$

Figure 6.2 shows that the relative univariate shrinkage factor of the BCTH data using $M_1 = 1$ as baseline and varying $M_2 = 0.1, 0.01, 0.001, 0.0001$ with 16000 cases and 16000 controls. It can be seen that the shrinkage factors with $M_2 = 0.1$ are very close to zero and so the posterior means are similar for both values of M. With $M_2 = 0.01$ a few effect sizes show a small amount of differential shrinkage but most of the posterior effect sizes do not change

that much. However, it can be seen that the shrinkage factors with $M_2 = 0.001$ start to move away from zero and there are now some shrinkage factors greater than a half. The extreme case of the shrinkage factors with $M_2 = 0.0001$ leads to many shrinkage factors more than a half and also a few close to 1 which implies posterior effect sizes very close to 0 with the M_2 prior variance. Figure 6.3 shows the relative univariate shrinkage factors of the BCTH data using $M_1 = 1$ as baseline and varying $M_2 = 0.1, 0.01, 0.001, 0.0001$ with 32000 cases and 32000 controls. It can be seen that they exhibit quite similar behaviour to the relative univariate shrinkage factors of the BCTH data with 16000 cases and 16000 controls. The values are generally closer to zero with the larger sample sizes as expected since the prior has less effect as the sample size increases. If we fix M_1 at 1 as in Chapter 4, both Figures 6.2 and 6.3 indicate that a suitable choice of M_2 is 0.001.

One might criticise the choice of the NG prior because the ratio between M_1, and M_2 is so large $(1/0.001 = 1000)$. In other words, the variance of the NG prior of the group with low FS scores was reduced 1000 times. Therefore, we suggest changing the baseline (M_1) from 1 to 0.01. The reason behind this is that the posterior means using $M_1 = 1$ and $M_2 = 0.01$ are similar. This should mean that the influence of the NG prior will be similar for $M_1 = 0.01$, and $M_1 = 1$.

We now check the relative shrinkage factors with the new baseline value of M_1. Figure 6.4(a) shows the relative univariate shrinkage factor calculated using $M_1 = 0.01$ and $M_2 = 0.001$ with 16000 cases and 16000 controls. It can be seen that more than two thirds of the effect sizes have SFs more than 0.1 and that the shrinkage factor does not exceed 0.5. Figure 6.4(b) shows the relative univariate shrinkage factor calculated using $M_1 = 0.01$ and $M_2 = 0.001$ with 32000 cases and 32000 controls. It can be seen that most of the shrinkage factors are less than 0.05 and do not exceed 0.3 rather similar to the $M_1 = 1$ case. The ratio between $M_1 = 0.01$ and $M_2 = 0.001$ is obviously much smaller than the ratio with baseline equal to 1. Here the variance for the NG prior of the group with low FS scores was reduced only 10 times (compared to 1000) which seems a more plausible reduction in prior variance.

Figure 6.5 shows the prior of β for $M = 0.01$ and $M = 0.001$ with $\lambda = 1/148$. It can be seen that as the M decreases the mass located close to zero increases.

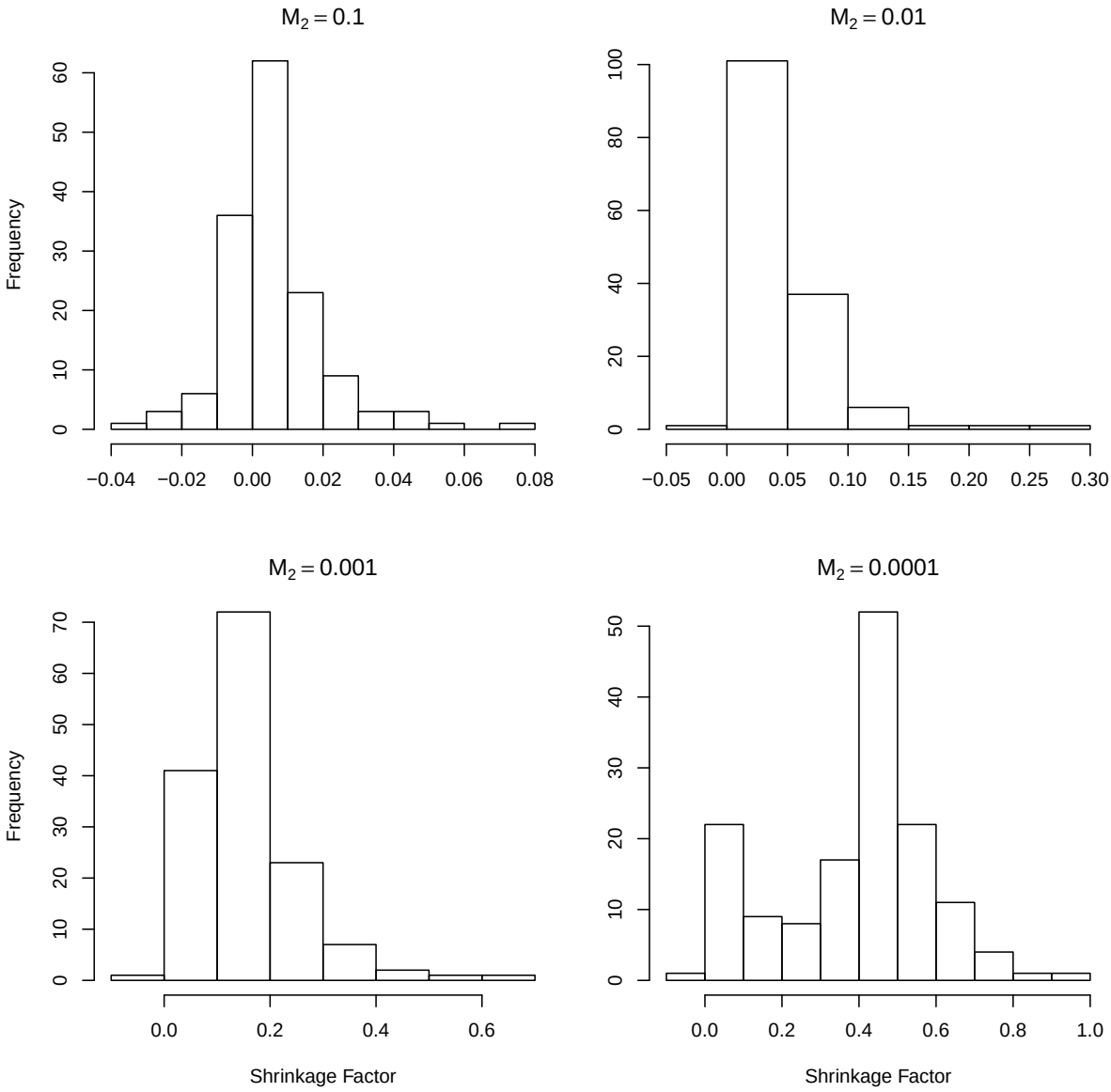


Figure 6.2: Histograms of the relative univariate shrinkage factors calculated using $M_1 = 1$ *and* $M_2 = 0.1, 0.01, 0.001, 0.0001$ *with 16000 cases and 16000 controls for the breast cancer top hits data.*

6.3 Full conditional distributions

In this section, we will calculate the full conditional distributions of each parameter. The full conditional distribution of a particular parameter is derived from the joint distribution of all the parameters and the data.

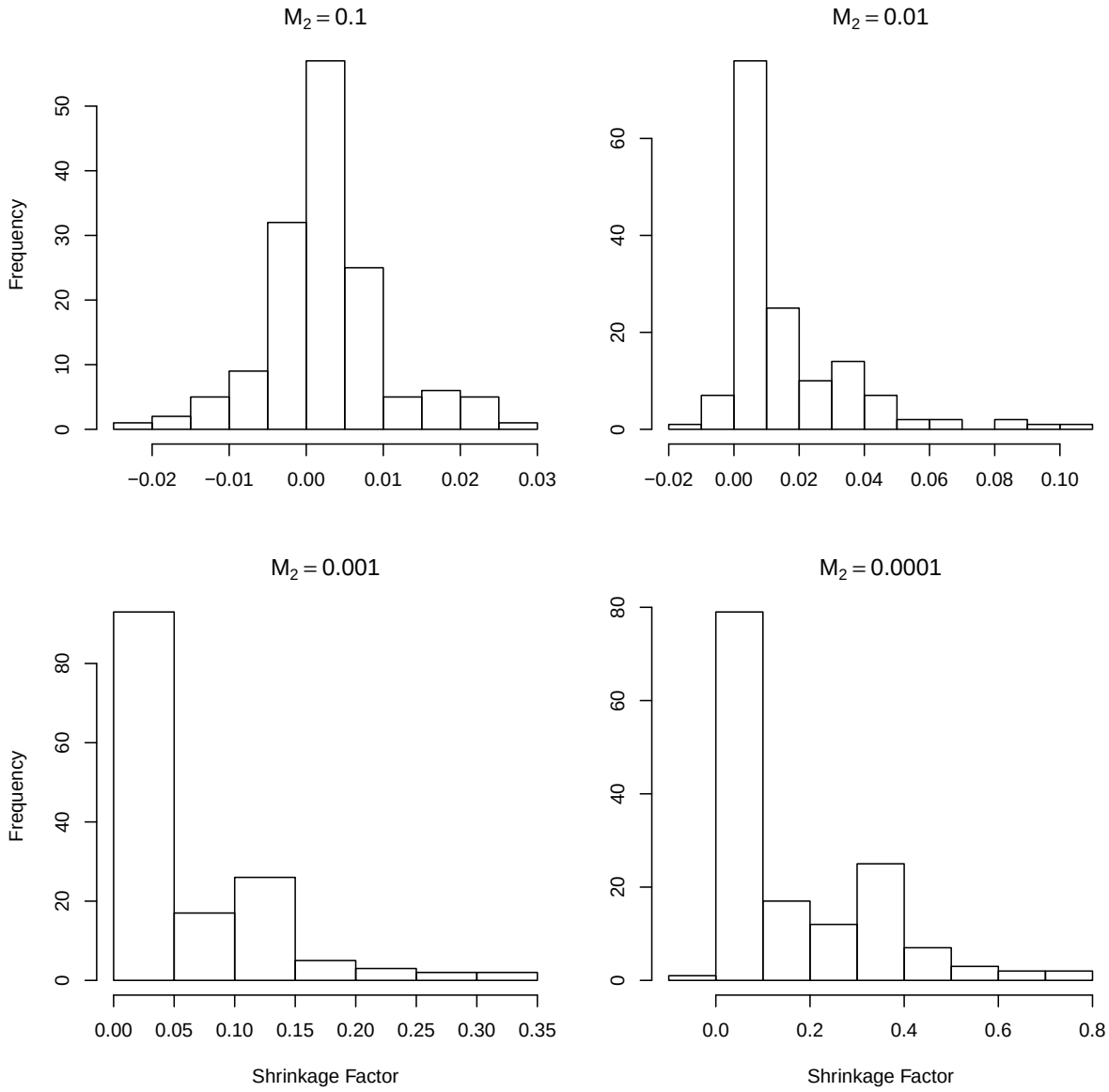
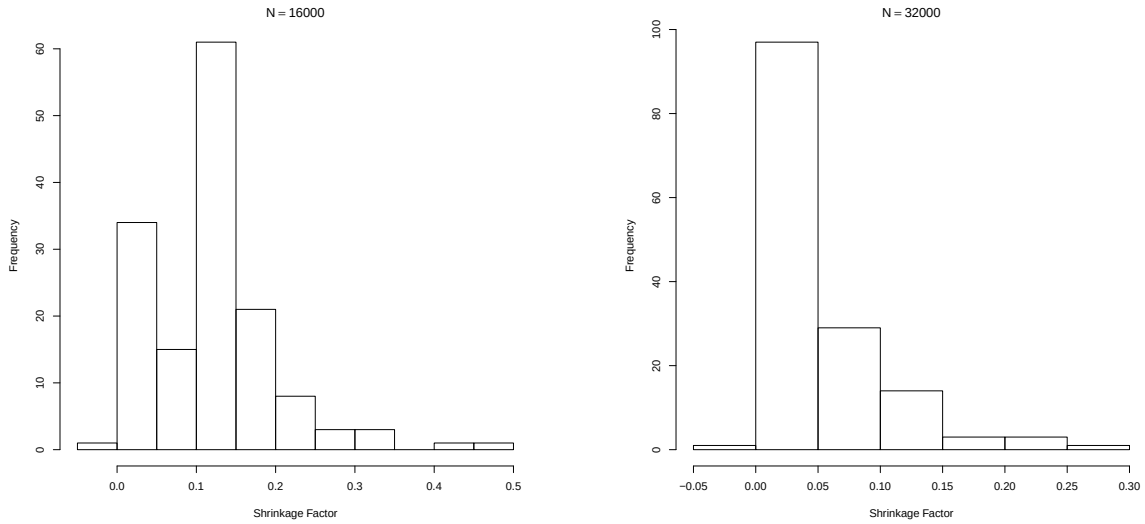


Figure 6.3: Histograms of the relative univariate shrinkage factors calculated using $M_1 = 1$ and $M_2 = 0.1, 0.01, 0.001, 0.0001$ with 32000 cases and 32000 controls for the breast cancer top hits data.

6.3.1 Full conditional distributions for α and $\boldsymbol{\beta}$

Here we use the same notation for the parameters as used in Chapter 3. The full conditional distribution for $\boldsymbol{\phi} = (\alpha, \boldsymbol{\beta})^T$ was calculated in Section 3.2.2 as

$$f(\boldsymbol{\phi} \mid \boldsymbol{\Lambda}, \lambda, \gamma^{-2}, \hat{\boldsymbol{\phi}}) \sim \boldsymbol{N}\left(\left(\boldsymbol{V}^{-1} + \boldsymbol{\Lambda}\right)^{-1} \boldsymbol{V}^{-1}\hat{\boldsymbol{\phi}}, \boldsymbol{V}^{-1} + \boldsymbol{\Lambda}\right), \tag{6.28}$$



(a) Applied on 16000 cases and 16000 controls.

(b) Applied on 32000 cases and 32000 controls.

Figure 6.4: Histograms for the relative univariate shrinkage factors calculated using $M_1 = 0.01$ *and* $M_2 = 0.001$ *with two different sample sizes.*

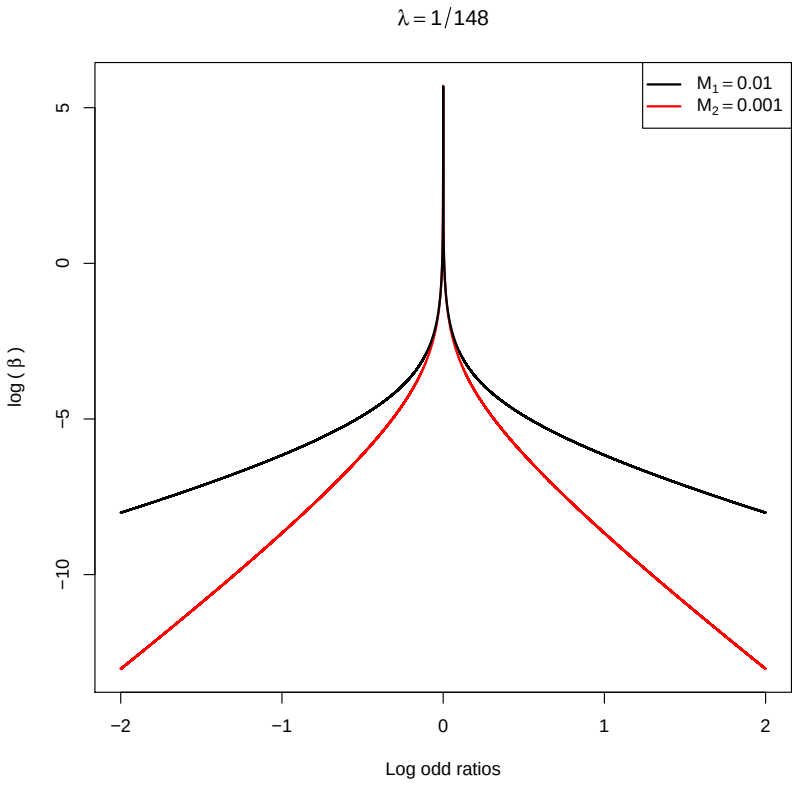


Figure 6.5: The prior $\pi(\beta)$ *for* $M_1 = 0.01$ *and* $M_2 = 0.001$ *with* $\lambda = 1/148$.

where $\mathbf{\Lambda} = \text{diag}\left(\frac{1}{0.1}, \frac{1}{\psi_{11}}, \ldots, \frac{1}{\psi_{p_1 1}}, \frac{1}{\psi_{12}}, \ldots, \frac{1}{\psi_{p_2 2}}, \frac{1}{\psi_{13}}, \ldots, \frac{1}{\psi_{p_3 3}}, \frac{1}{\psi_{14}}, \ldots, \frac{1}{\psi_{p_4 4}}, \right)$. The difference between the full conditional distributions for α and $\boldsymbol{\beta}$ in the standard NG prior and the modified NG prior, that includes the FS score its prior, is in the values of $\mathbf{\Lambda}$. In the standard NG prior, $\mathbf{\Lambda} = \text{diag}\left(\frac{1}{0.1}, \frac{1}{\psi_1}, \frac{1}{\psi_2}, \ldots, \frac{1}{\psi_p}\right)$, where each ψ is sampled from a distribution with the same value of λ and γ^{-2} for all SNPs. However, in the modified NG prior, each ψ_{ij} is sampled from one of four distributions with different values of λ_j and γ_j^{-2} based on the SNP's group $(1 \leq j \leq 4)$.

6.3.2 Full conditional distributions for ψ_{ij}

We calculate the full conditional distribution for ψ_{ij}, with λ_j and γ_j^{-2} determined by which FS score group (j) the SNP(i) belongs to. The likelihood does not contribute to the full conditional distribution of ψ_{ij}. The full conditional, can be derived from the joint distribution as follows

$$
\begin{aligned}
f\left(\psi_{ij} \mid \beta_{ij}, \lambda_j, \gamma_j^{-2}\right) &\propto \frac{1}{\sqrt{\psi_{ij}}} \exp\left(-\frac{\beta_{ij}^2}{2\psi_{ij}}\right) (\psi_{ij})^{\lambda_j - 1} \exp\left(-\frac{\psi_{ij}}{2\gamma_j^2}\right) && (6.29)\\
&= (\psi_{ij})^{-\frac{1}{2}} \exp\left(-\frac{\beta_{ij}^2}{2\psi_{ij}}\right) (\psi_{ij})^{\lambda_j - 1} \exp\left(-\frac{\psi_{ij}}{2\gamma_j^2}\right) && (6.30)\\
&= (\psi_{ij})^{\lambda_j - \frac{1}{2} - 1} \exp\left\{-\frac{1}{2}\left(\gamma_j^{-2}\psi_{ij} + \frac{\beta_{ij}^2}{2\psi_{ij}}\right)\right\} && (6.31)\\
&= (\psi_{ij})^{(\lambda_j - \frac{1}{2}) - 1} \exp\left\{-\frac{1}{2}\left(\gamma_j^{-2}\psi_{ij} + \frac{\beta_{ij}^2}{2\psi_{ij}}\right)\right\}. && (6.32)
\end{aligned}
$$

The expression is the kernel of a generalised inverse Gaussian (GIG) distribution with $m = \lambda_j - \frac{1}{2}, c = \gamma_j^{-2}$ and $d = \beta_{ij}^2$.

Full conditional distributions for λ_j

Here we will calculate the full conditional distribution for λ_j using the same approach as used in Section 3.2.4. The difference between the previous updating and this updating is that here we update λ_j for each group. As a result, the full conditional distribution of λ_j will be

extracted from the joint distribution as follows

$$\begin{aligned} f\left(\lambda_j \mid \beta_{ij}, \psi_{ij}, \gamma_j^{-2}, y_i, \boldsymbol{X}\right) &\propto \frac{\left(\frac{1}{2\gamma_j^2}\right)^{p_j\lambda_j}}{\left\{\Gamma\left(\lambda_j\right)\right\}^{p_j}} \times \left(\prod_{i=1}^{p_j}\psi_{ij}\right)^{\lambda_j-1} \times \prod_{i=1}^{p_j}\exp\left(-\frac{\psi_{ij}}{2\gamma_j^2}\right) \\ &\times \left(\gamma_j^{-2}\right)^{2-1} \times \left(\frac{M_k}{2\lambda_j}\right)^2 \times \exp\left(-\frac{M_k}{2\lambda_j}\gamma_j^{-2}\right) \\ &\times \pi\left(\lambda_j\right) \\ &= \frac{\left(\frac{1}{2\gamma_j^2}\right)^{p_j\lambda_j}}{\left\{\Gamma\left(\lambda_j\right)\right\}^{p_j}} \times \left(\prod_{i=1}^{p_j}\psi_{ij}\right)^{\lambda_j-1} \times \prod_{i=1}^{p_j}\exp\left(-\frac{\psi_{ij}}{2\gamma_j^2}\right) \\ &\times \left(\frac{M_k^2}{2}\right) \times \left(\frac{1}{2\lambda_j\gamma_j^2}\right) \times \left(\frac{1}{\lambda_j}\right) \times \exp\left(-\frac{M_k}{2\lambda_j}\gamma_j^{-2}\right) \\ &\times \pi\left(\lambda_j\right). \end{aligned}$$

Each λ_j is updated using Metropolis-Hasting approach. $\lambda_j' = \exp\left(\sigma_{\lambda_j}^2 z\right)\lambda_j$ is chosen to be the proposal distribution, where z is realisation of a standard normal distribution. Also, we choose an adjustment value for ${\gamma_j'}^2$ as follows

$${\gamma_j'}^2 = \frac{2\lambda_j\gamma_j^2}{2\lambda_j'}. \tag{6.33}$$

This trick is applied to ensure that the value of ${\gamma_j'}^2$ leads us to obtain the same variance of $\beta_{ij}|\lambda_j, \gamma_j^2$. After some algebra the acceptance probability of λ_j' can be shown to be

$$\min\left\{1, \frac{\pi\left(\lambda_j'\right)}{\pi\left(\lambda_j\right)} \frac{\left(\prod_{i=1}^{p_j}\psi_{ij}\right)^{\lambda_j'-1}}{\left(\prod_{i=1}^{p_j}\psi_{ij}\right)^{\lambda_j-1}} \frac{\left\{\Gamma\left(\lambda_j\right)\right\}^{p_j}}{\left\{\Gamma\left(\lambda_j'\right)\right\}^{p_j}} \frac{\prod_{i=1}^{p_j}\exp\left(-\frac{\psi_{ij}}{2\gamma_j'^2}\right)}{\prod_{i=1}^{p_j}\exp\left(-\frac{\psi_{ij}}{2\gamma_j'^2}\right)} \frac{\left(2\gamma_j^2\right)^{p_j\lambda_j}}{\left(2\gamma_j'^2\right)^{p_j\lambda_j}} \frac{\lambda_j'}{\lambda_j}\right\}. \tag{6.34}$$

The acceptance rate is controlled to be between $20\% - 30\%$ using the tuning parameter $\sigma_{\lambda_j}^2$.

6.3.3 Full conditional distributions for γ_j^{-2}

In this section, we will calculate the full conditional distribution for γ_j^{-2}. The likelihood does not contribute to the full conditional distribution of γ_j^{-2}. So, for $j = \{1, 3\}$ we derive it from

the joint distribution as follows

$$\begin{aligned} f\left(\gamma_j^{-2} \mid \beta_{ij}, \psi_{ij}, \lambda_j\right) &\propto \left(\gamma_j^{-2}\right)^{p_j\lambda_j} \times \exp\left(-\frac{\sum_{i=1}^{p_j}\psi_{ij}}{2}\gamma_j^{-2}\right) \times \left(\gamma_j^{-2}\right)^{2-1} \times \exp\left(-\frac{M_k}{2\lambda_j}\gamma_j^{-2}\right) \\ &= \left(\gamma_j^{-2}\right)^{(p_j\lambda_j+2)-1} \exp\left\{-\left(\frac{M_k}{2\lambda_j} + \frac{1}{2}\sum_{i=1}^{p_j}\psi_{ij}\right)\gamma_j^{-2}\right\}. \end{aligned} \tag{6.35}$$

This expression is the kernel of a gamma distribution with shape $e^* = p_j\lambda_j + 2$ and rate $f^* = \frac{M_k}{2\lambda_j} + \frac{1}{2}\sum_{i=1}^{p_j}\psi_{ij}$. This expression is used with the first group (FS > 0.5), where $M_1 = 0.01$ and the third group (FS < 0.5), where $M_2 = 0.001$.

However, for the second group ($j = 2$; FS $= 0.5$) the full conditional can be calculated as follows

$$\begin{aligned} f\left(\gamma_2^{-2} \mid \ldots\right) &= \left(\gamma_2^{-2}\right)^{p_2\lambda_2} \exp\left(-\frac{\sum_{i=1}^{p_2}\psi_{i2}}{2}\gamma_2^{-2}\right)\left\{wM_1^2\gamma_2^{-2}\exp\left(-\frac{M_1}{2\lambda_2}\gamma_2^{-2}\right)\right\} \\ &+ \left(\gamma_2^{-2}\right)^{p_2\lambda_2} \exp\left(-\frac{\sum_{i=1}^{p_2}\psi_{i2}}{2}\gamma_2^{-2}\right)\left\{(1-w)M_2^2\gamma_2^{-2}\exp\left(-\frac{M_2}{2\lambda_2}\gamma_2^{-2}\right)\right\} \\ &= \left\{wM_1^2\left(\gamma_2^{-2}\right)^{p_2\lambda_2+1}\exp\left[-\frac{1}{2}\left(\frac{M_1}{2\lambda_2} + \frac{\sum_{i=1}^{p_2}\psi_{i2}}{2}\right)\gamma_2^{-2}\right]\right\} \\ &+ \left\{(1-w)M_2^2\left(\gamma_2^{-2}\right)^{p_2\lambda_2+1}\exp\left[-\frac{1}{2}\left(\frac{M_2}{2\lambda_2} + \frac{\sum_{i=1}^{p_2}\psi_{i2}}{2}\right)\gamma_2^{-2}\right]\right\}. \\ \text{Let } A &= \frac{1}{2}\left(\frac{M_1}{\lambda_2} + \sum_{i=1}^{p_2}\psi_{i2}\right) \text{ and } B = \frac{1}{2}\left(\frac{M_2}{\lambda_2} + \sum_{i=1}^{p_2}\psi_{i2}\right) \\ f\left(\gamma_2^{-2} \mid \ldots\right) &= \left\{wM_1^2\left(\gamma_2^{-2}\right)^{p_2\lambda_2+1}\exp\left(A\gamma_2^{-2}\right)\right\} \\ &+ \left\{(1-w)M_2^2\left(\gamma_2^{-2}\right)^{p_2\lambda_2+1}\exp\left(B\gamma_2^{-2}\right)\right\} \qquad (6.36) \\ &= \left\{\frac{wM_1^2\Gamma\left(p_2\lambda_2+2\right)}{A^{p_2\lambda_2+2}}\left[\frac{A^{p_2\lambda_2+2}}{\Gamma\left(p_2\lambda_2+2\right)}\left(\gamma_2^{-2}\right)^{p_2\lambda_2+2-1}\exp\left(A\gamma_2^{-2}\right)\right]\right\} \\ &+ \left\{\frac{(1-w)M_2^2\Gamma\left(p_2\lambda_2+2\right)}{B^{p_2\lambda_2+2}}\left[\frac{B^{p_2\lambda_2+2}}{\Gamma\left(p_2\lambda_2+2\right)}\left(\gamma_2^{-2}\right)^{p_2\lambda_2+2-1}\exp\left(B\gamma_2^{-2}\right)\right]\right\} \qquad (6.37) \\ &\propto \left\{\frac{wM_1^2}{A^{p_2\lambda_2+2}}\left[\frac{A^{p_2\lambda_2+2}}{\Gamma\left(p_2\lambda_2+2\right)}\left(\gamma_2^{-2}\right)^{p_2\lambda_2+2-1}\exp\left(A\gamma_2^{-2}\right)\right]\right\} \\ &+ \left\{\frac{(1-w)M_2^2}{B^{p_2\lambda_2+2}}\left[\frac{B^{p_2\lambda_2+2}}{\Gamma\left(p_2\lambda_2+2\right)}\left(\gamma_2^{-2}\right)^{p_2\lambda_2+2-1}\exp\left(B\gamma_2^{-2}\right)\right]\right\} \qquad (6.38) \end{aligned}$$

$$\text{So } \gamma_2^{-2} \mid \ldots \sim \frac{wM_1^2}{A^{p_2\lambda_2+2}}\, Ga\left(p_2\lambda_2+2, A\right) + \frac{(1-w)\, M_2^2}{B^{p_2\lambda_2+2}}\, Ga\left(p_2\lambda_2+2, B\right) \tag{6.39}$$

$$= wM_1^2\, Ga\left(p_2\lambda_2+2, A\right) + (1-w)\, M_2^2 \left(\frac{A}{B}\right)^{p_2\lambda_2+2} Ga\left(p_2\lambda_2+2, B\right) \tag{6.40}$$

$$\sim \Delta_1\, Ga\left(p_2\lambda_2+2, A\right) + (1-\Delta_1)\; Ga\left(p_2\lambda_j+2, B\right), \tag{6.41}$$

where $\Delta_1 = \frac{wM_1^2}{wM_1^2 + \left(\frac{A}{B}\right)^{p_2\lambda_2+2}(1-w)M_2^2}$ and $f(\gamma_2^{-2} \mid \ldots)$ represents $f\left(\gamma_2^{-2} \mid \beta_{i2}, \psi_{i2}, \lambda_2, w\right)$.

In addition, for the fourth group with missing FS (FS = NA) the full conditional can be calculated as follows

$$\begin{aligned} f\left(\gamma_4^{-2} \mid \ldots\right) &= \left(\gamma_4^{-2}\right)^{p_4\lambda_4} \exp\left(-\frac{\sum_{i=1}^{p_4}\psi_{i4}}{2}\gamma_4^{-2}\right)\left\{hM_1^2\gamma_4^{-2}\exp\left(-\frac{M_1}{2\lambda_4}\gamma_4^{-2}\right)\right\} \\ &+ \left(\gamma_4^{-2}\right)^{p_4\lambda_4} \exp\left(-\frac{\sum_{i=1}^{p_4}\psi_{i4}}{2}\gamma_4^{-2}\right)\left\{(1-h)\, M_2^2\gamma_4^{-2}\exp\left(-\frac{M_2}{2\lambda_4}\gamma_4^{-2}\right)\right\} \\ &= \left\{hM_1^2\left(\gamma_4^{-2}\right)^{p_4\lambda_4+1}\exp\left[-\frac{1}{2}\left(\frac{M_1}{2\lambda_4}+\frac{\sum_{i=1}^{p_4}\psi_{i4}}{2}\right)\gamma_4^{-2}\right]\right\} \\ &+ \left\{(1-h)\, M_2^2\left(\gamma_4^{-2}\right)^{p_4\lambda_4+1}\exp\left[-\frac{1}{2}\left(\frac{M_2}{2\lambda_4}+\frac{\sum_{i=1}^{p_4}\psi_{i4}}{2}\right)\gamma_4^{-2}\right]\right\}. \end{aligned} \tag{6.42}$$

$$\text{Let } C = \frac{1}{2}\left(\frac{M_1}{\lambda_4}+\sum_{i=1}^{p_4}\psi_{i4}\right) \text{ and } D = \frac{1}{2}\left(\frac{M_2}{\lambda_4}+\sum_{i=1}^{p_4}\psi_{i4}\right)$$

$$\begin{aligned} f\left(\gamma_4^{-2} \mid \ldots\right) &= \left\{hM_1^2\left(\gamma_4^{-2}\right)^{p_4\lambda_4+1}\exp\left(C\gamma_4^{-2}\right)\right\} \\ &+ \left\{(1-h)\, M_2^2\left(\gamma_4^{-2}\right)^{p_4\lambda_4+1}\exp\left(D\gamma_4^{-2}\right)\right\} \end{aligned} \tag{6.43}$$

$$\begin{aligned} &= \left\{\frac{hM_1^2\Gamma\left(p_4\lambda_4+2\right)}{C^{p_4\lambda_4+2}}\left[\frac{C^{p_4\lambda_4+2}}{\Gamma\left(p_4\lambda_4+2\right)}\left(\gamma_4^{-2}\right)^{p_4\lambda_4+2-1}\exp\left(C\gamma_4^{-2}\right)\right]\right\} \\ &+ \left\{\frac{(1-h)\, M_2^2\Gamma\left(p_4\lambda_4+2\right)}{D^{p_4\lambda_4+2}}\left[\frac{D^{p_4\lambda_4+2}}{\Gamma\left(p_4\lambda_4+2\right)}\left(\gamma_4^{-2}\right)^{p_4\lambda_4+2-1}\exp\left(D\gamma_4^{-2}\right)\right]\right\} \end{aligned} \tag{6.44}$$

$$\begin{aligned} &\propto \left\{\frac{hM_1^2}{C^{p_4\lambda_4+2}}\left[\frac{C^{p_4\lambda_4+2}}{\Gamma\left(p_4\lambda_4+2\right)}\left(\gamma_4^{-2}\right)^{p_4\lambda_4+2-1}\exp\left(C\gamma_4^{-2}\right)\right]\right\} \\ &+ \left\{\frac{(1-h)\, M_2^2}{D^{p_4\lambda_4+2}}\left[\frac{D^{p_4\lambda_4+2}}{\Gamma\left(p_4\lambda_4+2\right)}\left(\gamma_4^{-2}\right)^{p_4\lambda_4+2-1}\exp\left(D\gamma_4^{-2}\right)\right]\right\} \end{aligned} \tag{6.45}$$

$$\text{So } \gamma_4^{-2} \mid \ldots \sim \frac{hM_1^2}{C^{p_4\lambda_4+2}}\, Ga\left(p_4\lambda_4+2, C\right) + \frac{(1-h)\, M_2^2}{D^{p_4\lambda_4+2}}\, Ga\left(p_4\lambda_4+2, D\right) \tag{6.46}$$

$$= hM_1^2\, Ga\left(p_4\lambda_4+2, C\right) + (1-h)\, M_2^2\left(\frac{C}{D}\right)^{p_4\lambda_4+2} Ga\left(p_4\lambda_4+2, D\right)$$

$$\sim \quad \Delta_2 \, Ga\left(p_4\lambda_4 + 2, C\right) + \left(1 - \Delta_2\right) \, Ga\left(p_4\lambda_4 + 2, D\right), \tag{6.47}$$

where $\Delta_2 = \frac{hM_1^2}{hM_1^2 + \left(\frac{C}{D}\right)^{p_4\lambda_4+2}(1-h)M_2^2}$ and $f\left(\gamma_4^{-2} \mid \ldots\right)$ represents $f\left(\gamma_4^{-2} \mid \beta_{i4}, \psi_{i4}, \lambda_4, h\right)$.

6.3.4 Full conditional distributions for w

In this section we will calculate the full conditional distribution for w. It can be seen that w occurs only in $\gamma_2^{-2} \mid \lambda_2, w$ and the prior on w so the full conditional distribution is given as follows. Let E represent the probability density given by $f_X\left(X = \gamma_2^{-2}\right)$, where $X \sim Ga\left(2, \frac{M_1}{2\lambda_2}\right)$ and F represent the probability density given by $f_Y\left(Y = \gamma_2^{-2}\right)$, where $Y \sim Ga\left(2, \frac{M_2}{2\lambda_2}\right)$. Then

$$\begin{aligned}
f\left(w \mid \lambda_2, \gamma_2^{-2}\right) &= \Big[wE + (1-w)\,F\Big]\pi\left(w\right) && (6.48)\\
&\propto \Big[wE + (1-w)\,F\Big] \times \Big[w^{2-1}\left(1-w\right)^{2-1}\Big] && (6.49)\\
&= w^{3-1}\left(1-w\right)^{2-1}E + w^{2-1}\left(1-w\right)^{3-1}F && (6.50)\\
&= \frac{\Gamma\left(3\right)\Gamma\left(2\right)}{\Gamma\left(5\right)}\Big[\frac{\Gamma\left(5\right)}{\Gamma\left(3\right)\Gamma\left(2\right)}w^{3-1}\left(1-w\right)^{2-1}\Big]E\\
&+ \frac{\Gamma\left(2\right)\Gamma\left(3\right)}{\Gamma\left(5\right)}\Big[\frac{\Gamma\left(5\right)}{\Gamma\left(2\right)\Gamma\left(3\right)}w^{2-1}\left(1-w\right)^{3-1}\Big]F. && (6.51)
\end{aligned}$$

It can be seen that $\left[\frac{\Gamma(5)}{\Gamma(3)\Gamma(2)}w^{3-1}\left(1-w\right)^{2-1}\right]$ is the probability density function of a Beta $(3,2)$ distribution and $\left[\frac{\Gamma(5)}{\Gamma(2)\Gamma(3)}w^{2-1}\left(1-w\right)^{3-1}\right]$ is the probability density function of a Beta $(2,3)$ distribution. Therefore, the full conditional distribution for w is given by

$$w \mid \lambda_2, \gamma_2^{-2} \quad \sim \quad \Delta_3 \text{ Beta}\left(3,2\right) + \left(1-\Delta_3\right)\text{Beta}\left(2,3\right), \tag{6.52}$$

where $\Delta_3 = \frac{\frac{\Gamma(3)\Gamma(2)}{\Gamma(5)}E}{\frac{\Gamma(3)\Gamma(2)}{\Gamma(5)}E + \frac{\Gamma(2)\Gamma(3)}{\Gamma(5)}F} = \frac{E}{E+F}$.

6.3.5 Full conditional distributions for h

In this section we will calculate the full conditional distribution for h. It can be seen that h occurs only in $\gamma_4^{-2} \mid \lambda_4, h$ and the prior on h so the full conditional distribution is given as follows. Let G represent the probability density given by $f_X\left(X = \gamma_4^{-2}\right)$, where $X \sim$

$Ga\left(2, \frac{M_1}{2\lambda_4}\right)$ and H represent the probability density given by $f_Y\left(Y = \gamma_4^{-2}\right)$, where $Y \sim Ga\left(2, \frac{M_2}{2\lambda_4}\right)$. Then

$$\begin{aligned} f\left(h \mid \lambda_4, \gamma_4^{-2}\right) &= \Big[hG + (1-h)\,H\Big]\pi\left(h\right) && (6.53)\\ &\propto \Big[hG + (1-h)\,H\Big] \times \Big[h^{1-1}\left(1-h\right)^{4-1}\Big] && (6.54)\\ &= h^{2-1}\left(1-h\right)^{4-1}G + h^{1-1}\left(1-h\right)^{5-1}H && (6.55)\\ &= \frac{\Gamma\left(2\right)\Gamma\left(4\right)}{\Gamma\left(6\right)}\Big[\frac{\Gamma\left(6\right)}{\Gamma\left(2\right)\Gamma\left(4\right)}h^{2-1}\left(1-h\right)^{4-1}\Big]G \\ &+ \frac{\Gamma\left(1\right)\Gamma\left(5\right)}{\Gamma\left(6\right)}\Big[\frac{\Gamma\left(6\right)}{\Gamma\left(1\right)\Gamma\left(5\right)}h^{1-1}\left(1-h\right)^{5-1}\Big]H. && (6.56)\end{aligned}$$

It can be seen that $\left[\frac{\Gamma(6)}{\Gamma(2)\Gamma(4)}h^{2-1}\left(1-h\right)^{4-1}\right]$ is the probability density function of a Beta $(2,4)$ distribution and $\left[\frac{\Gamma(6)}{\Gamma(1)\Gamma(5)}h^{1-1}\left(1-h\right)^{5-1}\right]$ is the probability density function of a Beta $(1,5)$ distribution. Therefore, the full conditional distribution for h is given by

$$h \mid \lambda_4, \gamma_4^{-2} \sim \Delta_4 \text{ Beta}\left(2,4\right) + \left(1-\Delta_4\right)\text{Beta}\left(1,5\right), \quad (6.57)$$

where $\Delta_4 = \frac{\frac{\Gamma(2)\Gamma(4)}{\Gamma(6)}G}{\frac{\Gamma(2)\Gamma(4)}{\Gamma(6)}G + \frac{\Gamma(1)\Gamma(5)}{\Gamma(6)}H} = \frac{\Gamma(2)\Gamma(4)G}{\Gamma(2)\Gamma(4)G+\Gamma(1)\Gamma(5)H} = \frac{\Gamma(4)G}{\Gamma(4)G+\Gamma(5)H} = \frac{G}{G+4H}$.

Chapter 7

The effect of incorporating FS scores into the effect size prior on simulated data

In Chapter 6 we discussed incorporating FS scores into the NG prior. The FS scores were divided into four groups: FS > 0.5, FS $= 0.5$, FS < 0.5, and missing FS (FS $=$ NA). Each group was given a different prior based on the potential deleterious nature of the SNPs in each group as determined by the FS score. In addition, the full conditional distributions were calculated for all parameters in the model.

Throughout this chapter we will compare only the standard NG prior (without incorporating FS scores into the prior) and the modified NG prior (with incorporating FS scores into the prior) and we will discuss the effect of incorporating FS scores into the prior for the effect size on the simulated data in our 8 scenarios (see Tables 4.1 and 4.2). Particularly, we will discuss in detail the effect of incorporating FS scores into the prior for the effect size in Scenario 4 which has a moderate level of LD.

7.1 The standard Normal-Gamma prior and the Modified Normal-Gamma prior

In this section we will compare the performance of the standard NG prior and modified NG prior via ROC curves in all 8 Scenarios (see Tables 4.1 and 4.2). For the standard NG prior,

we set the expectation of the variance of the effect size to be 0.01. For the modified NG prior, we set $M_1 = 0.01$ and $M_2 = 0.001$ as discussed in Section 6.2.

7.1.1 Placing the common and rare causal SNPs in the same FS scores group

In this section we will discuss the effect of incorporating the FS scores into the prior for effect size when we place the common and rare causal SNPs in the same FS score group. Here we are considering the ROC curves for FPR < 0.1 to assess the performance using ROC curves. To partition the SNPs into four groups, we suggested two scenarios: randomly sample FS scores from their distribution in the iCOGs data, in which the proportions for the four groups are $0.02, 0.05, 0.3, 0.61$ respectively (realistic proportions) and randomly sample FS scores with the proportions for the four groups: $0.17, 0.17, 0.17, 0.49$ respectively (less realistic proportion because the proportion of FS ≥ 0.5 SNPs is perhaps too high). The reason for choosing two sets of proportions is to assess the effect of the number of SNPs in the SNP group considered most likely to be deleterious. We considered all four scenarios for setting the group of the common and rare causal SNPs: both common and rare causal SNPs into Group 1 (FS > 0.5), both common and rare causal SNPs into Group 2 (FS $= 0.5$), both common and rare causal SNPs into Group 3 (FS < 0.5), and both common and rare causal SNPs into Group 4 (FS $=$ NA).

Figures 7.1 and 7.2 (FPR < 0.1) show the ROC curves for the eight scenarios in Tables 4.1 and 4.2 for the standard NG prior and the modified NG prior with the common and rare causal SNPs in the same group where the proportions for the four FS score groups are $0.02, 0.05, 0.3, 0.61$. It can be seen that the performance of the modified NG prior is better than the performance of the standard NG regardless of which group contains the common and rare causal SNP group. Moreover, it can be noticed that the performance of the modified NG setting both the common and rare causal SNP into Group 1 and Group 2 is almost always better than the performance of the modified NG setting both the common and rare causal SNP into Group 3 and Group 4. However, Figure 7.2(c) shows that the performance of the modified NG setting both the common and rare causal SNP into Group 3 perform better than the

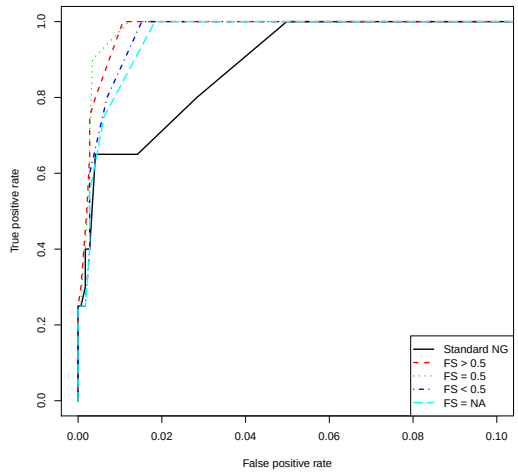
others. The reason for this is not clear. It seems unlikely to be the high LD level because this observation cannot be made for Scenario 4 in spite of the fact that it has a moderate level of LD.

Figures 7.3 and 7.4 $(\text{FPR} < 0.1)$ show the ROC curves for the standard NG prior and the modified NG prior setting the common and rare causal SNPs in the same group where the proportions for the four FS score groups being $0.17, 0.17, 0.17, 0.49$. It can be seen that the performance of the modified NG prior setting both the common and rare causal SNP into Group 1 and Group 2 is again better than the performance of the standard NG except in Scenario 7 where the performance of the standard NG is the best (see Figure 7.4(c)). Moreover, it can be noticed that the performance of the modified NG setting both the common and rare causal SNP into Group 3 and Group 4 is comparable with the performance of the standard NG (except in Scenario 7).

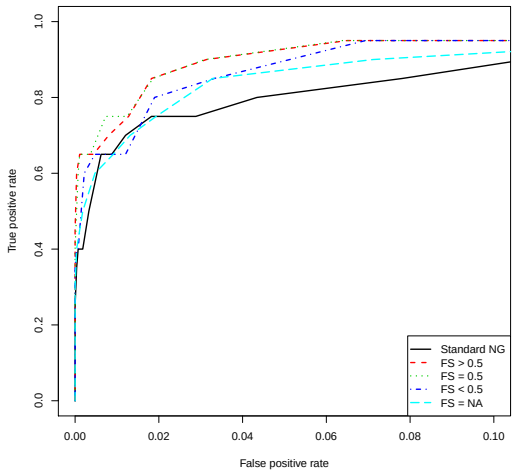
It seems that the proportion of SNPs in each group could affect the performance of the modified NG as seen in Figures 7.1-7.4, although it seems that the performance of the modified NG setting both the common and rare causal SNP into Group 1 or Group 2 is generally better than the others for both sets of the proportion of SNPs in each group considered. It is also of interest, since it is probably more likely, to consider the effect of placing the two causal SNPs into different FS score groups. To avoid considering too many scenarios we focus on just one. For the rest of this chapter, we focus on the scenario with moderate LD level. Thus, we will discuss in the next sections the effect of incorporating FS scores into the NG prior in Scenario 4 (see Table 4.1 and Figures 7.1(d) and 7.3(d)) in detail. The proportions for the four groups 1 to 4 of $0.02, 0.05, 0.3,$ and 0.61 respectively will be considered in the next sections.

7.1.2 Placing the common and rare causal SNPs into different FS score groups in Scenario 4

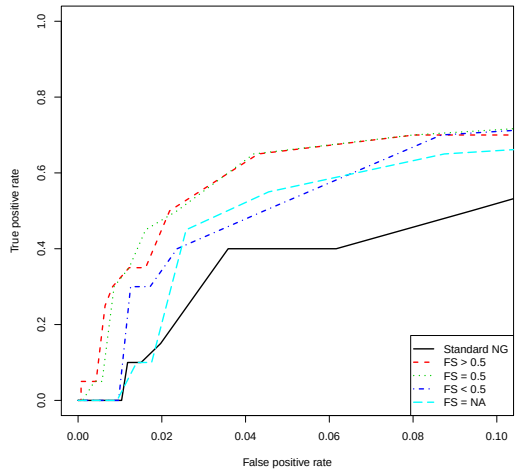
We will discuss the effect of incorporating the FS scores into the prior for effect size in Scenario 4 by setting the common and rare causal SNP into different groups. We will compare this performance with the performance of the standard NG and the performance where both of them are set into the same FS score group.



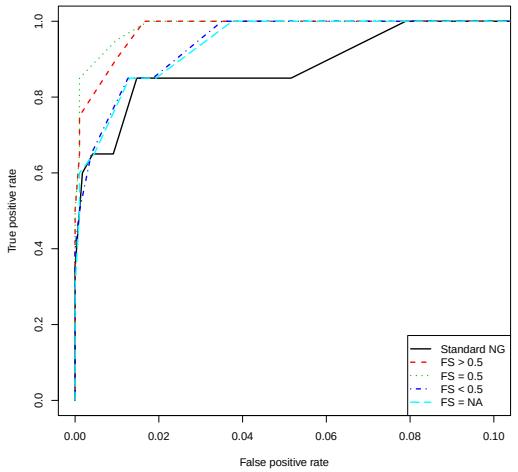
(a) ROC curves for the standard NG prior and the modified NG prior applied to Scenario 1. The number of SNPs in 1 to 4 the groups are 8, 15, 89, 179 respectively.



(b) ROC curves for the standard NG prior and the modified NG prior applied to Scenario 2. The number of SNPs in 1 to 4 the groups are 8, 15, 87, 166 respectively.

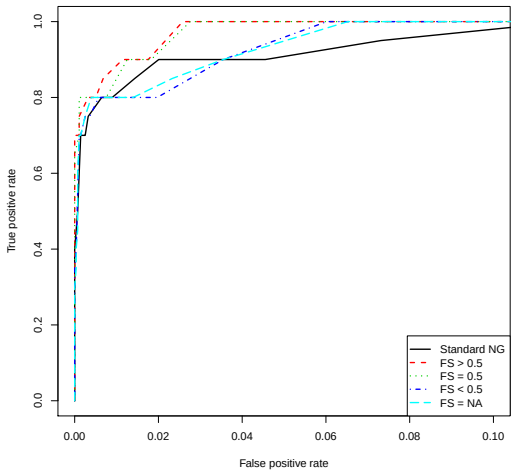


(c) ROC curves for the standard NG prior and the modified NG prior applied to Scenario 3. The number of SNPs in 1 to 4 the groups are 8, 15, 89, 169 respectively.

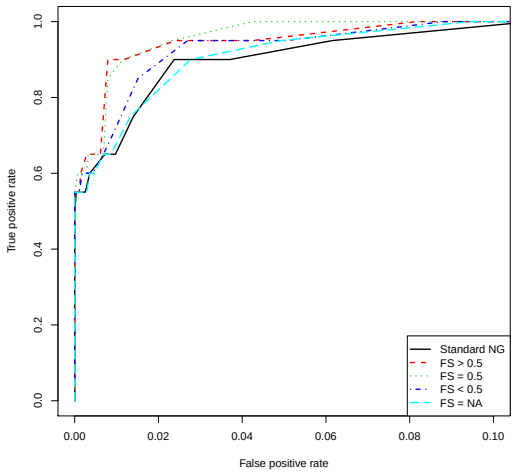


(d) ROC curves for the standard NG prior and the modified NG prior applied to Scenario 4. The number of SNPs in 1 to 4 the groups are 7, 15, 89, 176 respectively.

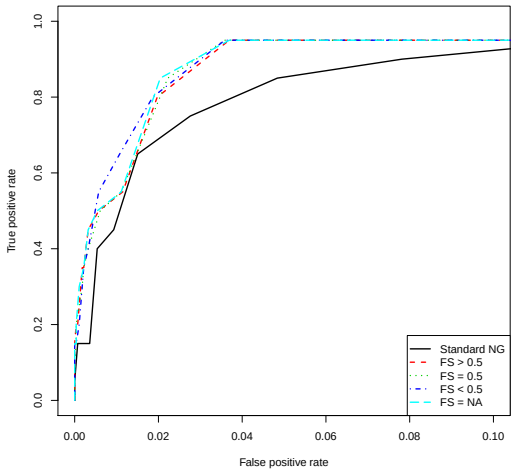
Figure 7.1: ROC curves for the credible interval approach using the standard NG prior $(M = 0.01)$*, and the modified NG prior* $(M_1 = 0.01$ *and* $M_2 = 0.001)$ *with: both causal SNPs in Group1* $(FS > 0.5)$*; both causal SNPs in Group2* $(FS = 0.5)$*; both causal SNPs in Group3* $(FS < 0.5)$*; and both causal SNPs in Group4* $(FS = NA)$*. The methods are applied to 10 simulated datasets from Hapgen2 with two causal SNPs having fixing odds ratios of 1.08 and 1.13 and having different MAFs (see Table 4.1). Each dataset has 16000 cases and 16000 controls. The MCMC was run for 20,000 iterations with 2,000 burn-in and thinning by 50. The proportions of SNPs in each of groups 1 to 4 are 0.02, 0.05, 0.3, 0.61 respectively.*



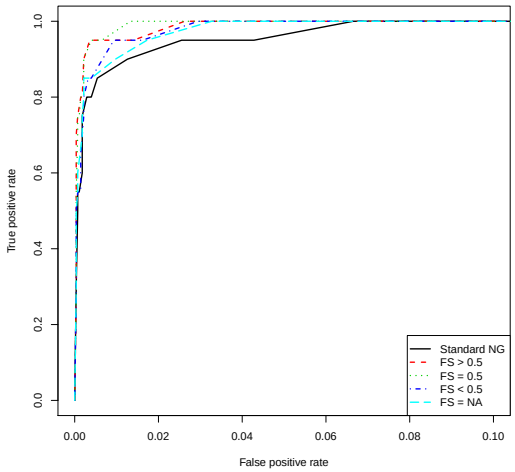
(a) ROC curves for the standard NG prior and the modified NG prior applied to Scenario 5. The number of SNPs in 1 to 4 the groups are 8, 15, 88, 170 respectively.



(b) ROC curves for the standard NG prior and the modified NG prior applied to Scenario 6. The number of SNPs in 1 to 4 the groups are 8, 15, 88, 169 respectively.

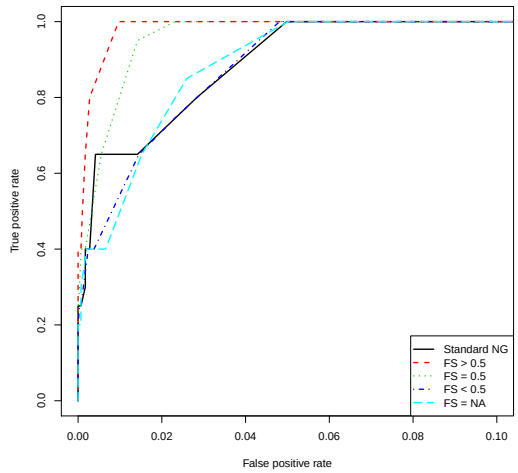


(c) ROC curves for the standard NG prior and the modified NG prior applied to Scenario 7. The number of SNPs in 1 to 4 the groups are 7, 15, 89, 170 respectively.

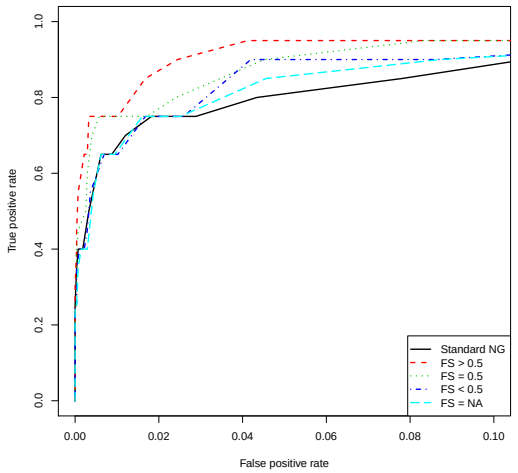


(d) ROC curves for the standard NG prior and the modified NG prior applied to Scenario 8. The number of SNPs in 1 to 4 the groups are 7, 15, 88, 170 respectively.

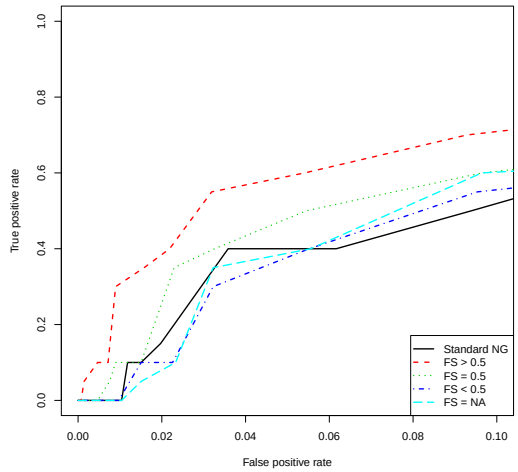
Figure 7.2: ROC curves for the credible interval approach using the standard NG prior $(M = 0.01)$*, and the modified NG prior* $(M_1 = 0.01$ *and* $M_2 = 0.001)$ *with: both causal SNPs in Group1* $(FS > 0.5)$*; both causal SNPs in Group2* $(FS = 0.5)$*; both causal SNPs in Group3* $(FS < 0.5)$*; and both causal SNPs in Group4* $(FS = NA)$*. The methods are applied to 10 simulated datasets from Hapgen2 with two causal SNPs having fixing odds ratios of 1.08 and 1.13 and having different MAFs (see Table 4.1). Each dataset has 32000 cases and 32000 controls. The MCMC was run for 20,000 iterations with 2,000 burn-in and thinning by 50. The proportions of SNPs in each of groups 1 to 4 are 0.02, 0.05, 0.3, 0.61 respectively.*



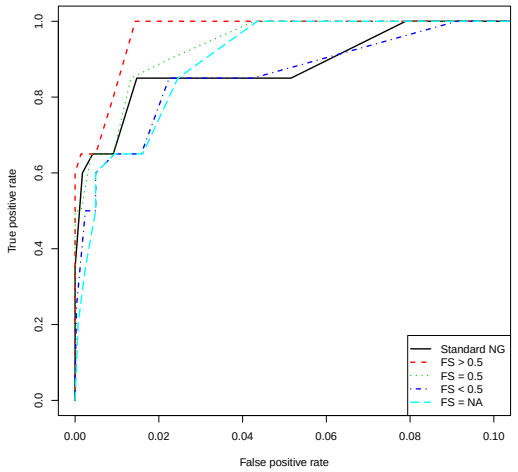
(a) ROC curves for the standard NG prior and the modified NG prior applied to Scenario 1. The number of SNPs in 1 to 4 the groups are 52, 50, 49, 140 respectively.



(b) ROC curves for the standard NG prior and the modified NG prior applied to Scenario 2. The number of SNPs in 1 to 4 the groups are 46, 47, 50, 133 respectively.

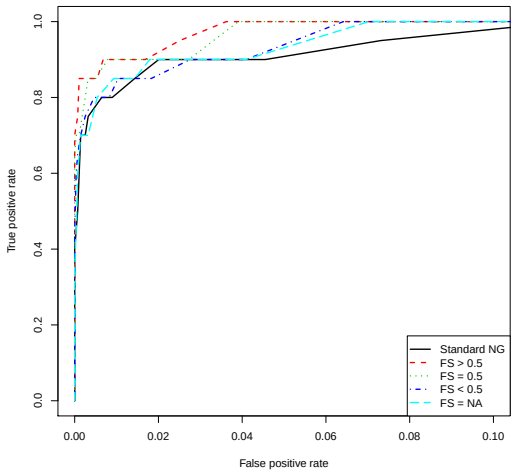


(c) ROC curves for the standard NG prior and the modified NG prior applied to Scenario 3. The number of SNPs in 1 to 4 the groups are 48, 49, 50, 134 respectively.

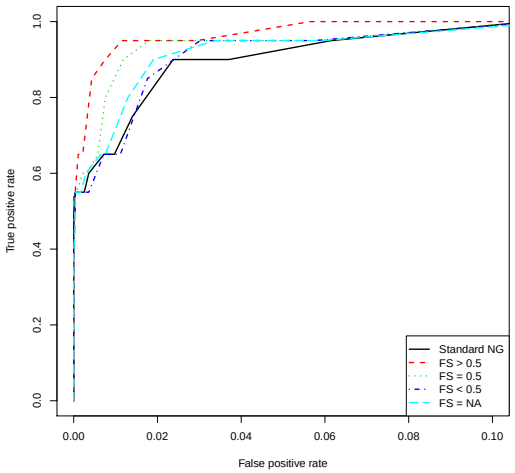


(d) ROC curves for the standard NG prior and the modified NG prior applied to Scenario 4. The number of SNPs in 1 to 4 the groups are 50, 50, 49, 138 respectively.

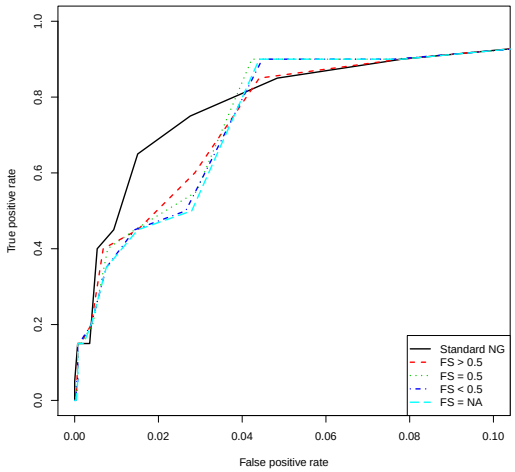
Figure 7.3: ROC curves for the credible interval approach using the standard NG prior $(M = 0.01)$*, and the modified NG prior* $(M_1 = 0.01$ *and* $M_2 = 0.001)$ *with: both causal SNPs in Group1* $(FS > 0.5)$*; both causal SNPs in Group2* $(FS = 0.5)$*; both causal SNPs in Group3* $(FS < 0.5)$*; and both causal SNPs in Group4* $(FS = NA)$*. The methods are applied to 10 simulated datasets from Hapgen2 with two causal SNPs having fixing odds ratios of 1.08 and 1.13 and having different MAFs (see Table 4.1). Each dataset has 16000 cases and 16000 controls. The MCMC was run for 20,000 iterations with 2,000 burn-in and thinning by 50. The proportions of SNPs in each of groups 1 to 4 are 0.17, 0.17, 0.17, 0.49 respectively.*



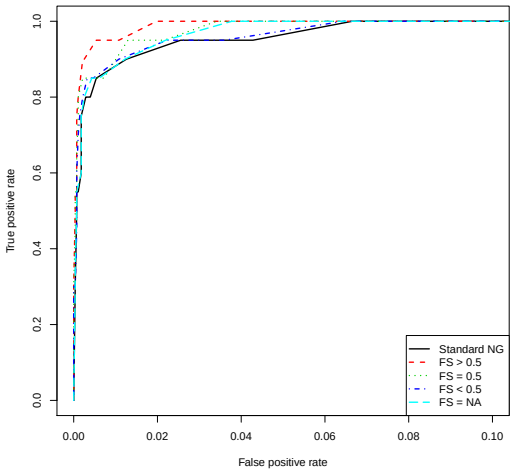
(a) ROC curves for the standard NG prior and the modified NG prior applied to Scenario 5. The number of SNPs in 1 to 4 the groups are 48, 49, 50, 134 respectively.



(b) ROC curves for the standard NG prior and the modified NG prior applied to Scenario 6. The number of SNPs in 1 to 4 the groups are 47, 49, 50, 134 respectively.



(c) ROC curves for the standard NG prior and the modified NG prior applied to Scenario 7. The number of SNPs in 1 to 4 the groups are 48, 49, 49, 135 respectively.



(d) ROC curves for the standard NG prior and the modified NG prior applied to Scenario 8. The number of SNPs in 1 to 4 the groups are 48, 49, 48, 135 respectively.

Figure 7.4: ROC curves for the credible interval approach using the standard NG prior $(M = 0.01)$*, and the modified NG prior* $(M_1 = 0.01$ *and* $M_2 = 0.001)$ *with: both causal SNPs in Group1* $(FS > 0.5)$*; both causal SNPs in Group2* $(FS = 0.5)$*; both causal SNPs in Group3* $(FS < 0.5)$*; and both causal SNPs in Group4* $(FS = NA)$*. The methods are applied to 10 simulated datasets from Hapgen2 with two causal SNPs having fixing odds ratios of 1.08 and 1.13 and having different MAFs (see Table 4.1). Each dataset has 32000 cases and 32000 controls. The MCMC was run for 20,000 iterations with 2,000 burn-in and thinning by 50. The proportions of SNPs in each of groups 1 to 4 are 0.17, 0.17, 0.17, 0.49 respectively.*

Scenarios	Common Causal SNP	Rare Causal SNP
Scenario 1 (S 1)	Group 1 (FS > 0.5)	Group 4 (FS $= NA$)
Scenario 2 (S 2)	Group 2 (FS $= 0.5$)	Group 4 (FS $= NA$)
Scenario 3 (S 3)	Group 4 (FS $= NA$)	Group 1 (FS > 0.5)
Scenario 4 (S 4)	Group 3 (FS < 0.5)	Group 4 (FS $= NA$)

Table 7.1: The four Scenarios of placing the common and rare causal SNPs into different FS scores group.

We suggest four scenarios for setting the common and rare causal SNPs. In all four scenarios one SNP is placed in Group 4 (FS $= NA$) because it is highly likely to obtain a causal SNP with an unknown FS scores. Table 7.1 shows the four scenarios chosen: Scenario 1 (S 1) sets the common causal SNP into Group 1 (FS > 0.5) and the rare causal SNP into Group 4 (FS $= NA$). This scenario was chosen because a causal SNP has a high FS score and the other causal SNP is likely to have a missing FS score. Scenario 2 (S 2) sets the common causal SNP in Group 2 (FS $= 0.5$) and the rare causal SNP in Group 4 (FS $= NA$). This scenario was chosen to investigate how the performance is affected when the common causal SNP placed in a group with lower prior variance compared to the variance of Group 1. Scenario 3 (S 3) sets the common causal SNP into Group 4 (FS $= NA$) and the rare causal SNP into Group 1 (FS > 0.5). This scenario was chosen to reverse the causal SNPs group of Scenario 1. Scenario 4 (S 4) sets the common causal SNP into Group 3 (FS < 0.5) and the rare causal SNP into Group 4 (FS $= NA$). This scenario was chosen to represent a worst case likely to be encountered.

Figure 7.5 shows ROC curves for Scenarios S1-S4. Each ROC curve also Shows the performance of the modified NG when both causal SNPs are placed in either of the two FS score groups for that particular scenario. The standard NG is also shown.

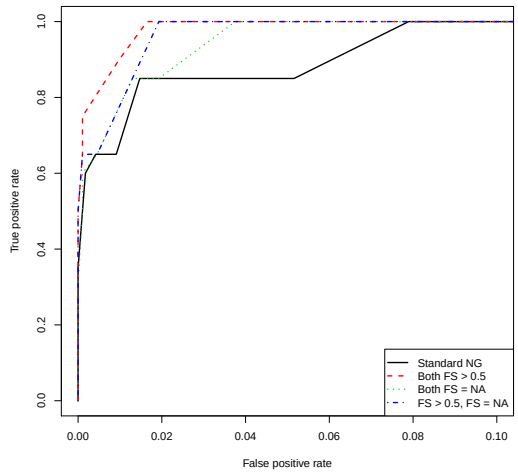
Figure 7.5 shows that the performance of the standard NG is the worse than the performance of the modified NG regardless of where the common and rare causal SNPs are located. Within the modified NG analyses, the performance of the modified NG setting both SNPs

into Group 1 (FS > 0.5) or Group 2 (FS $= 0.5$) is the best but it performs relatively poorly if both are set into Group 4 (FS $= NA$). In addition, it can be seen that when the two causal SNPs are separated into two groups, the performance is in between the performance of the modified NG setting both causal SNPs into Group 1 (FS > 0.5) or Group 2 (FS $= 0.5$) and the performance of the modified NG setting both causal SNPs into Group 4 (FS $= NA$).

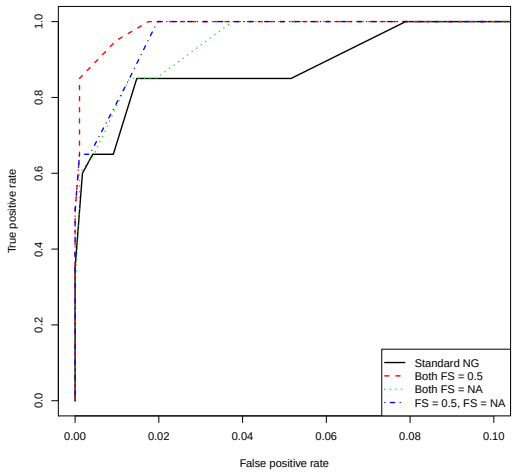
Since we are using credible intervals to select SNPs it is of interest to compare the maximum CI size that detects the causal SNPs in Scenarios S1-S4. Detection here means that zero is not in the interval. Table 7.2 shows the mean maximum detection credible interval size (across 10 datasets) for the common and rare causal SNPs for the four scenarios of separating causal SNPs into two different groups (S 1 - S 4). The mean maximum detection CI size for the common causal SNP when the common causal SNP is placed into Group 1 is 41.5%, Group 2 is 40.9%, Group 3 is 37.2%, and Group 4 is 36.4%. It can be seen that the mean maximum detection CI size of the common causal SNP increases as the SNP is placed in the groups with higher FS scores (S 1 and S 2). In addition, it can be noticed that the mean maximum detection CI size of the rare causal SNP increases dramatically when the SNP is placed in Group 1 (75.1%) compared to its mean maximum detection CI size where the SNP is put Group 4 (60.3%). Setting the common causal SNP into Group 3 (FS < 0.5; S 3) and setting the rare causal SNP into Group 4 (FS $= NA$; S 4) is the worst scenario as might be expected.

7.1.3 Setting SNPs within the LD block into different FS score groups

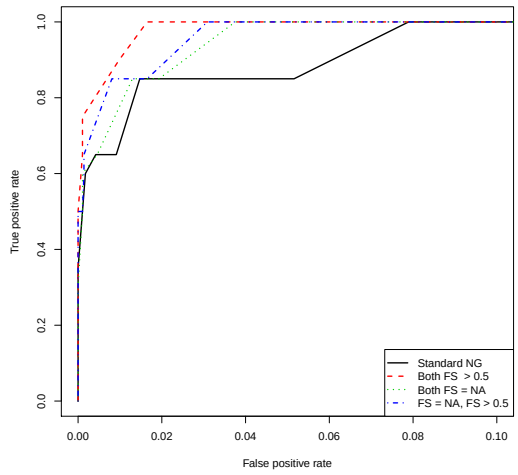
The results in Table 4.7 indicated that, in some situations, the NG prior approach is unable to pick out the causal SNP when there are SNPs present that are highly correlated with it. To investigate how incorporating functional information can improve the causal SNP detection we consider placing the causal SNP and other highly correlated SNPs into different FS scores groups. In a more systematic way in Scenario 4. See Table 4.7 for a descrption of the LD structure of the Scenario 4 LD block. Specifically we vary the FS score group of SNP46, SNP47, and SNP65. We suggest three scenarios for SNPs within the LD block containing the common causal SNP. Table 7.3 shows the three scenarios chosen: LD 1 sets the common causal SNP into Group 1 (FS > 0.5) and sets all SNPs within the LD block into Group 4



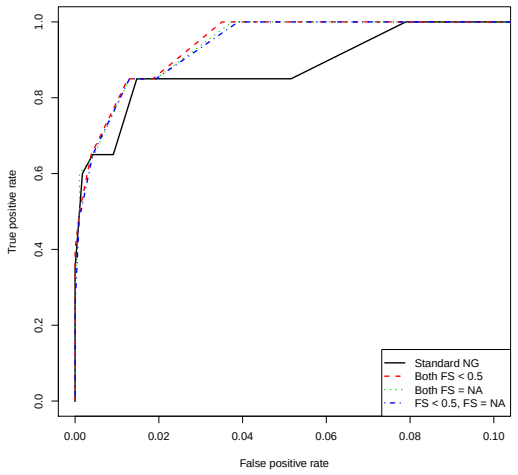
(a) ROC curves for the standard NG, the modified NG with both causal SNPs in Group 1, the modified NG with both causal SNPs in Group 4, the modified NG with the common causal SNP in Group 1 and the rare causal SNP in Group 4. $M_1 = 0.01$, and $M_2 = 0.001$. The proportions for the 4 groups are $6, 15, 89, 177$ respectively.



(b) ROC curves for the standard NG, the modified NG with both causal SNPs in Group 2, the modified NG with both causal SNPs in Group 4, the modified NG with the common causal SNP in Group 2 and the rare causal SNP in Group 4. $M_1 = 0.01$, and $M_2 = 0.001$. The proportions for the 4 groups are $5, 16, 89, 177$ respectively.



(c) ROC curves for the standard NG, the modified NG with both causal SNPs in Group 1, the modified NG with both causal SNPs in Group 4, the modified NG with the common causal SNP in Group 4 and the rare causal SNP in Group 1. The proportions for the 4 groups are $6, 15, 89, 177$ respectively.



(d) ROC curves for the standard NG, the modified NG with both causal SNPs in Group 3, the modified NG with both causal SNPs in Group 4, the modified NG with the common causal SNP in Group 3 and the rare causal SNP in Group 4. $M_1 = 0.01$, and $M_2 = 0.001$. The proportions for the 4 groups are $5, 15, 90, 177$ respectively.

Figure 7.5: ROC curves for the standard NG and modified NG prior where the two causal SNPs are either placed in the same FS scores group or in different FS scores groups.

SNPs	**46**	**212**
Mean CI size S 1	41.5%	60.3%
Mean CI size S 2	40.9%	60.5%
Mean CI size S 3	37.2%	75.1%
Mean CI size S 4	36.4%	60.1%

Table 7.2: The mean maximum detection credible interval of the common and the rare causal SNPs for Scenario 4 in Table 4.1. The maximum detection credible interval is calculated using $1 - 2 \times min\left\{Pr\left(\beta \mid \hat{\beta}, V\right) > 0, Pr\left(\beta \mid \hat{\beta}, V\right) < 0\right\}$. *SNP46 in blue is the common causal SNP and SNP212 in red is the rare causal SNP.*

(FS $= NA$). This scenario was chosen to represent the best scenario that could be chosen. LD 2 sets the common causal SNP (SNP46), SNP47, and SNP65 into Group 1 (FS > 0.5) and places the other SNPs within the LD block into Group 4 (FS $= NA$). This scenario was chosen to investigate the influence of combining the causal SNP with the SNPs most correlated with it in the same group. LD 3 places the SNP most correlated with the common causal SNP (SNP47) into Group 1 (FS > 0.5) and sets the other SNPs within the LD block (including the common causal SNP (SNP46)) into Group 4 (FS $= NA$). This scenario was chosen to represent the worst case scenario where a highly correlated SNP is placed in a high prior variance FS score group. Note the rare causal SNP is set into Group 3 (FS < 0.5) for all three scenarios. Here we will focus only on the SNPs within the LD block containing the common causal SNP. Particularly, we will concentrate on the common causal SNP (SNP46) and the SNP with the highest LD with the common causal SNP (SNP47), with $r^2 = 0.9$.

Table 7.4 shows the mean maximum detection credible interval size (across 10 datasets) for the standard NG, and the three scenarios LD1, LD2, and LD 3. It can be seen that the mean maximum detection CI size of the common causal SNP (SNP46) in the standard NG and LD 1 scenario are similar at 43.1% and 43.2% respectively. This is because the variance of the NG prior for the common causal SNP in both scenarios is 0.01. However, the mean maximum

Scenarios	SNP46	SNP47	SNP65	Other correlated SNP
LD 1	Group 1 (> 0.5)	Group 4 ($= NA$)	Group 4 ($= NA$)	Group 4 ($= NA$)
LD 2	Group 1 (> 0.5)	Group 1 (> 0.5)	Group 1 (> 0.5)	Group 4 ($= NA$)
LD 3	Group 4 ($= NA$)	Group 1 (> 0.5)	Group 4 ($= NA$)	Group 4 ($= NA$)

Table 7.3: The three Scenarios placing the common causal SNP (SNP46), the two SNPs highly correlated with it (SNP47 and SNP65), and the other correlated SNPs in the LD block (see Table 4.7) into different FS score groups.

SNPs	25	30	34	**46**	47	48	58	65	71	**212**
r^2	0.83	0.81	0.83	**1**	0.9	0.84	0.8	0.82	0.82	**0.09**
Mean CI size NG	7.2%	13.4%	12.4%	43.1%	35.8%	14.6%	5.1%	8%	7.7%	68.2%
Mean CI size LD 1	8.1%	11.1%	10.9%	43.2%	27.9%	11.5%	4.8%	3.4%	3.4%	62.1%
Mean CI size LD 2	7.1%	10%	9.8%	38.7%	33.6%	10.1%	3.6%	4.9%	4.1%	62.6%
Mean CI size LD 3	7.9%	11.5%	11.2%	33.7%	37.5%	11.3%	3.9%	4%	3.7%	62.7%

Table 7.4: The mean maximum detection credible interval is calculated using $1 - 2 \times min\left\{Pr\left(\beta \mid \hat{\beta}, V\right) > 0, Pr\left(\beta \mid \hat{\beta}, V\right) < 0\right\}$. r^2 *with the common causal SNP for SNPs in its LD block for Scenario 4 in Table 4.1, where the LD block is defined as* $r^2 \geq 0.8$. *SNP46 in blue is the common causal SNP and SNP212 in red is the rare causal SNP.*

detection CI sizes of the common causal SNP in LD 2 and LD 3 scenarios are reduced to 38.7% and 33.7% respectively. The reason behind this is that in LD 2, the common causal SNP is located in the same group as SNP47, the SNP most correlated with it, and in LD 3 SNP47 is located in Group 1 whereas the causal SNP is located in Group 4.

For the SNP most correlated with the common SNP (SNP47), it can be noticed that the mean maximum detection CI sizes (across 10 datasets) is reduced to 27.9% in the LD 1 scenario and 33.6% in the LD 2 scenario compared with a mean maximum detection CI size for the standard NG of 35.8%. Therefore, it is more likely that the causal SNP is the SNP selected within the LD block especially in the LD 1. In the LD 3 scenario, however, the mean

maximum detection CI sizes of SNP47 is 37.5% slightly higher than the size for the causal SNP, but only marginally. Therefore, it looks as though even if the causal SNP is placed in the lowest FS score group and the SNP most correlated with it is placed in the highest FS scores group, the modified NG still does not shrink the posterior effect sizes of the causal SNP too drastically. The mean maximum detection CI sizes of the rare causal SNP in the three scenarios are decreased to 62.1%, 62.6%, and 62.7% respectively compared with the mean CI size of the standard NG of 68.2%

Actual selection of SNPs within the LD block across datasets

In Section 4.9.2 we compared four statistical methods (NG, HL, SLR, and PiMASS) in terms of SNP selection within the LD block in Scenario 4 (see Table 4.1). The NG prior was applied with the prior variance of the effect size set to 1. However, here we will compare the performance of the standard NG prior with the performance of the modified NG in the three scenarios where SNPs within the LD block containing the common causal SNP are placed in various FS scores groups: (LD 1, LD, 2, and LD 3) in terms of SNP selection. Note, the variance of the NG prior for the standard NG prior is now 0.01. Also $M_1 = 0.01$ and $M_2 = 0.001$ (see Section 6.2).

Table 7.5 shows the total number of times a SNP was selected out of the 10 datasets in Scenario 4. For the standard NG, 25 SNPs were selected at 50% CI, and 25, 26, 25 SNPs were selected for the LD scenarios 1 to 3 respectively at 40% CI. These CI sizes were selected to ensure the total number of SNPs selected are almost the same.

There are three SNPs that are selected it at least one dataset in all three LD scenarios but not the standard NG (SNP35, SNP129, and SNP142). This is somewhat surprising since SNP35 and SNP142 are in Group 3 and Group 4 respectively. Also SNP143 is selected in some datasets in LD 2 and LD 3 but not in LD 1 and the standard NG. Plus both SNP44 and SNP227 were selected by the standard NG but not in any dataset for LD 1, LD 2, or LD 3. Therefore, the effect of the modified NG is very different for SNP44 and SNP142 even though they are both in Group 4 (see Table 7.5).

It can be seen that the standard NG and all three LD scenarios detect the rare causal SNP (SNP212) in 10 out of the 10 datasets although it was placed in Group 3. The maximum CI

SNPs	r^2 with SNP46	r^2 with SNP212	MAF	FS scores group	NG (50% CI)	NG-(40% CI)-LD 1	NG-(40% CI)-LD 2	NG-(40% CI)-LD 3
SNP35	0.12	0.18	0.05	3	0	3	3	3
SNP44	0.01	0	0.01	4	4	0	0	0
SNP46	1	0.09	0.31	1, 1, 4	3	5	3	2
SNP47	0.9	0.08	0.33	4, 1, 1	4	3	3	4
SNP129	0.03	0	0.01	1	0	2	4	3
SNP142	0.01	0.15	0.39	4	0	2	2	2
SNP143	0.01	0.18	0.37	4	0	0	1	1
SNP212	0.09	1	0.10	3	10	10	10	10
SNP227	0.01	0.03	0.21	3	4	0	0	0
Total number of SNPs selected					25	25	26	25

Table 7.5: Total number of times a SNP was selected out of the 10 datasets in Scenario 4 using 50% *credible intervals in the standard NG,* 40% *credible intervals in the modified NG using FS scores for three different scenarios (LD 1, LD 2, and LD 3).*

size to detect the rare causal SNP is 70% in the standard NG and approximately 65% in all three LD scenarios. In addition, it can be noticed that the common causal SNP (SNP46) is selected by the standard NG in 3 out of the 10 datasets, LD 1 in 5 out of the 10 datasets, LD 2 in 3 out of the 10 datasets, and LD 3 in 2 out of the 10 datasets.

The maximum CI size to select the common causal SNP is 55% in the standard NG, 55% in LD 1, 50% in LD 2, and 45% in LD 3. Furthermore, it can be seen that the standard NG detects SNP47, the most correlated SNP with the common causal SNP, in 4 out of the 10 datasets, LD 1 selects SNP47 in 3 out of the datasets, LD 2 detects the SNP in 3 out of the 10 datasets, and LD 3 selects the SNP in 4 out of the 10 datasets. The maximum CI size to select SNP47 is 60% in the standard NG, 45% in LD 1, 60% in LD 2, and 65% in LD 3.

It seems that the modified NG approach in LD 1 improves localisation of the signal within the LD block. In the standard NG the common causal SNP (SNP46) was selected in 3 out of the 10 datasets and (SNP47) was selected in 4 out of the 10 datasets, whereas in the LD 1 scenario, the modified NG selects the common causal SNP (SNP46) in 5 out of the 10 datasets and SNP47 was selected in 3 out of the 10 datasets. However, in the LD 2 and LD 3 scenarios,

the modified NG selects SNP46 in 3 and 2 out of the 10 datasets respectively and SNP47 in 3 and 4 datasets respectively. Therefore, the performance is less good in these more challenging scenarios.

It is of interest to investigate the shrinkage of SNPs in each of the FS scores groups in some specific datasets. SNP46 is always selected in the first dataset, and SNP47 is always selected in the seventh, eighth, and ninth datasets in all scenarios. Also both SNPs are selected together in the standard NG in the third datasets. Therefore, we will discuss the shrinkage factor for the three LD scenarios (LD 1, LD 2, and LD 3) specifically in the first, third, and eighth datasets.

Shrinkage factor for the three LD scenarios

In Section 7.1.3 we indicated the selection of SNPs within the LD block by incorporating the FS scores into the NG prior and applied this on three LD scenarios (LD 1, LD 2, and LD 3). Here, we will discuss the shrinkage factor for the three LD scenarios (LD 1, LD 2, and LD 3) generally using three datasets (first, third, and eighth dataset). Moreover, we will discuss in detail the amount of shrinkage for two SNPs of interest: the common causal SNP (SNP46), and the most correlated SNP with the common causal SNP (SNP47). This might help to explain the performance in each scenario.

The shrinkage factor (SF) can be calculated as follows

$$\text{SF} = 1 - \frac{\mathbb{E}\left(\beta \mid \hat{\beta}\right)}{\hat{\beta}}, \tag{7.1}$$

where $\mathbb{E}\left(\beta \mid \hat{\beta}\right)$ represent the mean posterior distribution of the effect size and $\hat{\beta}$ represents the Maximum Likelihood Estimate (MLE) of the effect size from multiple logistic regression.

Posterior mean densities

The posterior mean effect sizes is a term which is required to calculate the SF. As a result, the posterior mean densities for the effect sizes for the four groups will be shown for Scenario 4. Figures 7.6(a), 7.6(c), and 7.6(e) show the estimated densities of the posterior means of the

effect size for each of the four groups of SNPs in LD 1 using the first, third, and eighth dataset respectively. There are not enough observation in Group 1 to estimate the density. So we just plot the values individually. Generally, it can be seen that many of the posterior means for the effect sizes in Group 1 and Group 2 are away from zero (except in the third dataset (see Figure 7.6(c)) where most are close to zero), but most of those placed in Group 4 are close to zero. Group 3, (that includes the rare causal SNP) posterior means are mostly near to zero but there are small number away from zero. The common causal SNP (SNP46) is located in Group 1 and its posterior mean in the three datasets is $-0.034, -0.042, -0.022$ respectively. SNP47 which is highly correlated with SNP46 is located in Group 4 and its posterior mean in the three datasets is $0, -0.013, -0.021$ respectively (see Table 7.6).

Figures 7.7(a), 7.7(c), and 7.7(e) show the posterior mean densities (or actual posterior means) for the effect size of the four groups of SNPs in LD 2 using the first, third, and eighth dataset respectively. Moreover, Figures 7.8(a), 7.8(c), and 7.8(e) show the posterior mean densities (or actual posterior means) for the effect size of the four groups of SNPs in LD 3 using the first, third, and eighth dataset respectively. It can be seen that these Figures show much the same pattern as Figures 7.6(a), 7.6(c), and 7.6(e).

Shrinkage factors by FS scores group

Shrinkage factors (SF) were calculated using Equation 7.1 for the three LD scenarios (LD 1, LD 2, and LD 3) for the first, third, and eighth dataset.

Figures 7.6(b), 7.7(b), and 7.8(b) show the SF for the effect sizes for the three LD scenarios in the first dataset. It can be seen that they have almost the same in pattern that the majority of SNPs located in Group 3 and Group 4 were shrunk more than the SNPs located in Group 1 and Group 2. However, there are a few SNPs located in Group 3 and Group 4 that are shrunk less than the SNPs located in Group 1 or Group 2. For SNPs in Group 1, 4 out of the 6 SNPs have SF < 0.9. In addition, there are some SNPs with extreme SF in all three LD scenarios (not shown in the Figures). For example, in LD 1 there are six SNPs (SNP123, and SNP124 located in Group 2 and SNP33, SNP143, SNP165, and SNP178 located in Group 4) that have a SF larger than 1.2 or less than 0.7. The highest SF is 17.7 (SNP143) and the lowest SF is -15.5 (SNP124). These estimate SFs mostly occur as a result of instability in MLEs caused

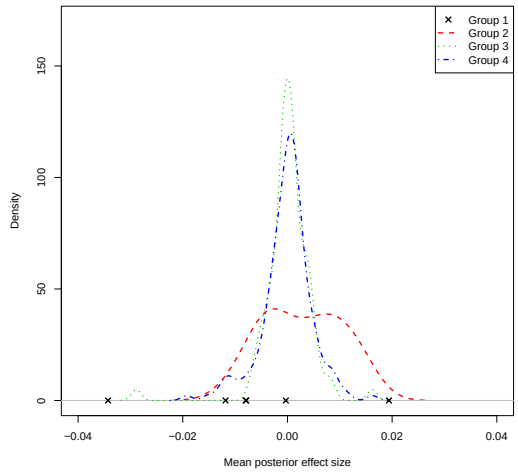
by high correlation of the SNPs.

Figures 7.6(d), 7.7(d), and 7.8(d) show the SF for the effect sizes for three LD scenarios in the third dataset. In addition, Figures 7.6(f), 7.7(f), and 7.8(f) show the SF for the effect sizes for three LD scenarios in the eighth dataset. The pattern in these Figures are similar to those in Figures 7.6(b), 7.7(b), and 7.8(b).

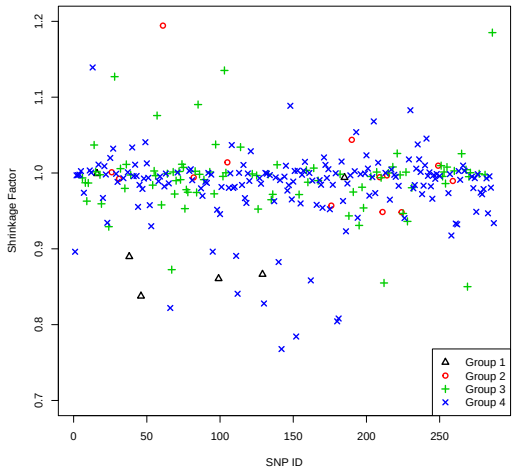
Shrinkage factors of the common causal SNP and the SNP most correlated with it across datasets

Table 7.6 illustrates the differences in shrinkage factors between the common causal SNP and the SNP most highly correlated with it (SNP46, and SNP47 respectively). This table shows the MLE of the effect size of both SNPs for 10 datasets, the posterior mean of the effect size and the shrinkage factor (SF) for three different LD scenarios (see Section 7.1.3).

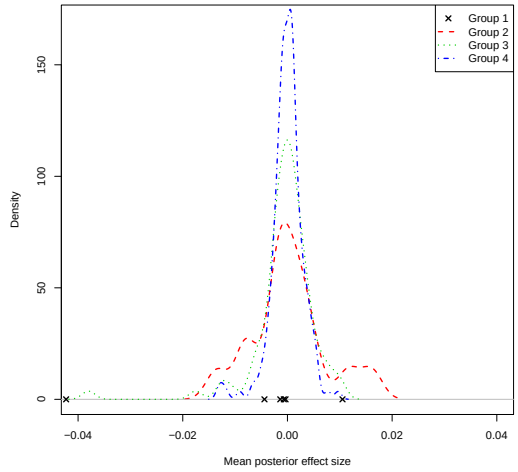
It can be seen that shrinkage factors for SNP46 are generally closer to 0 than those of SNP47. This is much more apparent in LD 1 and LD 2. Moreover, the posterior means for SNP47 have a wider range of values than for SNP46. It can be noticed that there is a lot of variation by dataset. Surprisingly there is relatively little variation across the LD 1, LD 2, and LD 3 scenarios.



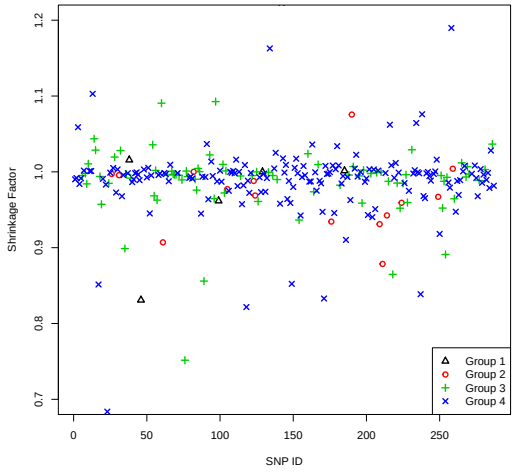
(a) Posterior mean densities for groups 2-4 and actual posterior means for Group 1 for dataset 1.



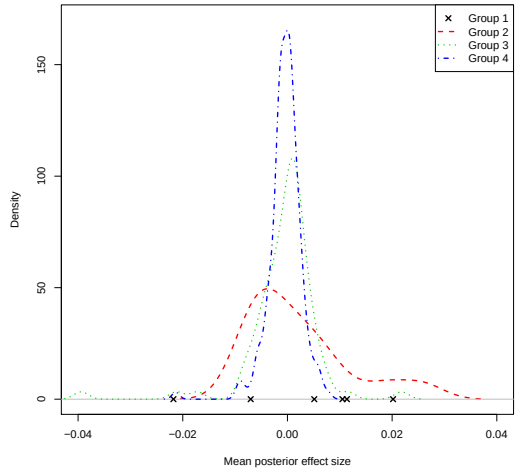
(b) Shrinkage factors for the 4 groups using dataset 1.



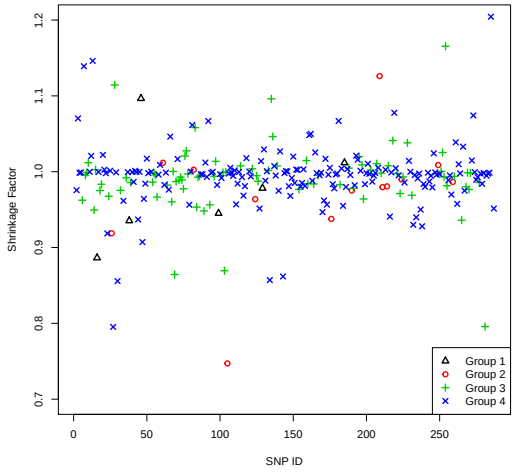
(c) Posterior mean densities for groups 2-4 and actual posterior means for Group 1 for dataset 3.



(d) Shrinkage factors for the 4 groups using dataset 3.

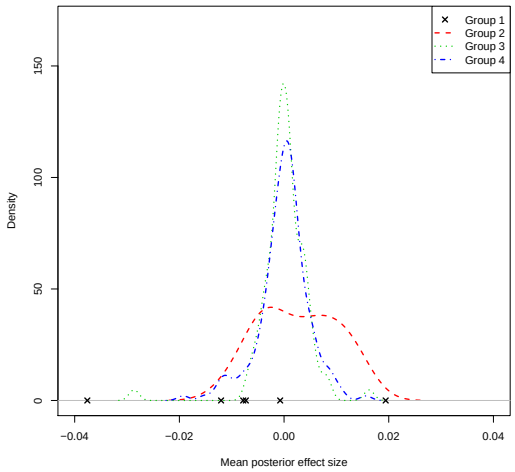


(e) Posterior mean densities for groups 2-4 and actual posterior means for Group 1 for dataset 8.

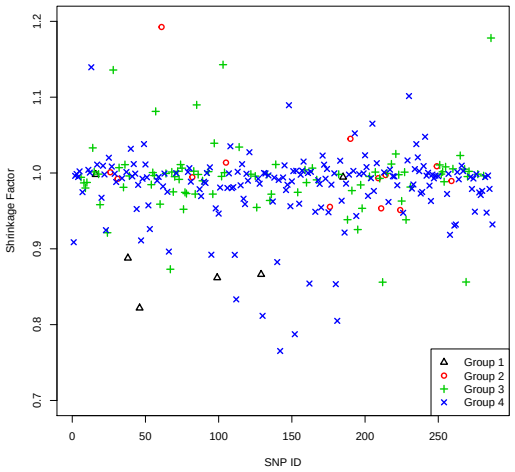


(f) Shrinkage factors for the 4 groups using dataset 8.

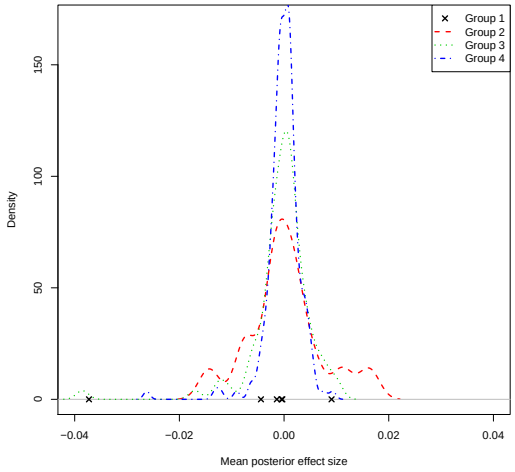
Figure 7.6: Posterior mean densities (or actual values) and shrinkage factors for Scenario 4 for LD1 with 16000 cases and 16000 controls.



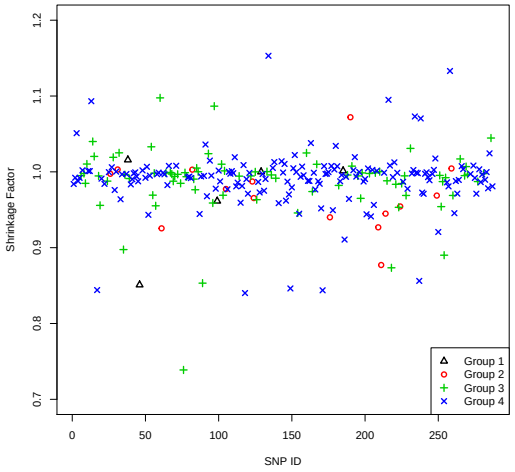
(a) Posterior mean densities for groups 2-4 and actual posterior means for Group 1 for dataset 1.



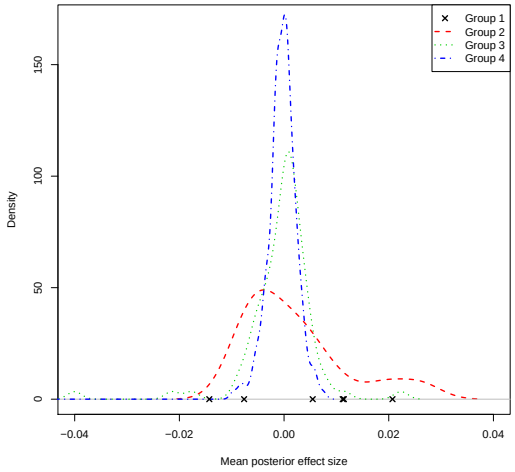
(b) Shrinkage factors for the 4 groups using dataset 1.



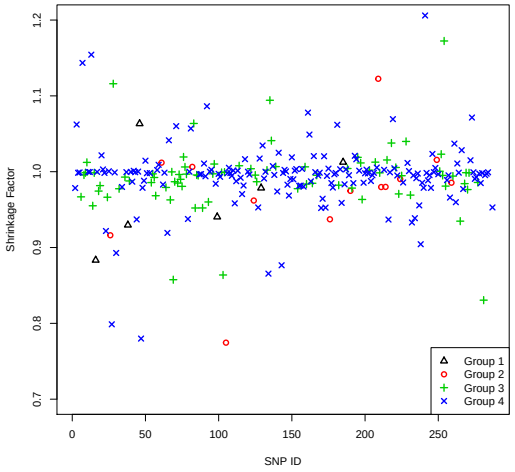
(c) Posterior mean densities for groups 2-4 and actual posterior means for Group 1 for dataset 3.



(d) Shrinkage factors for the 4 groups using dataset 3.

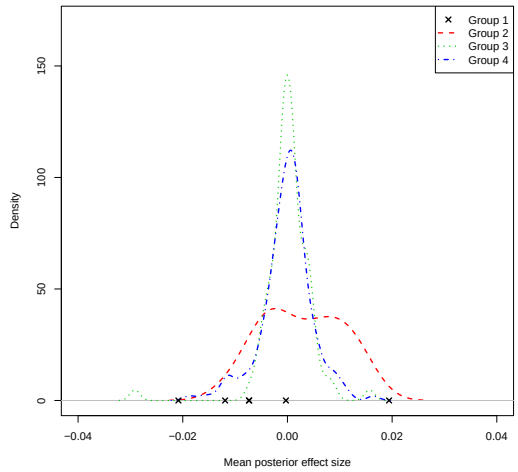


(e) Posterior mean densities for groups 2-4 and actual posterior means for Group 1 for dataset 8.

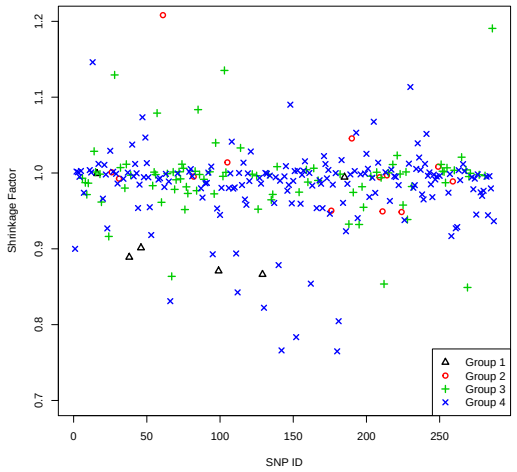


(f) Shrinkage factors for the 4 groups using dataset 8.

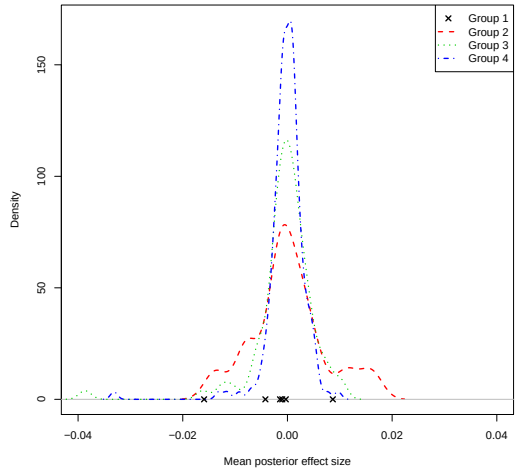
Figure 7.7: Posterior mean densities (or actual values) and shrinkage factors for Scenario 4 for LD2 with 16000 cases and 16000 controls.



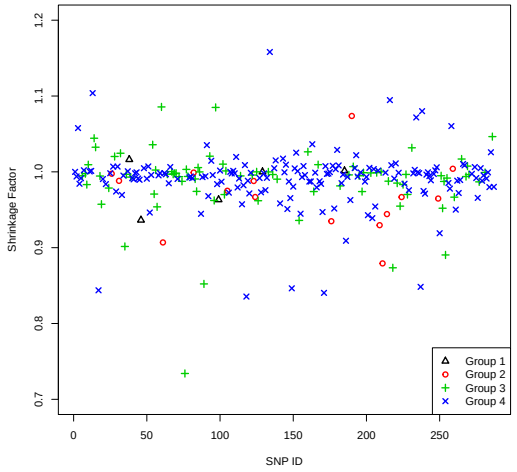
(a) Posterior mean densities for groups 2-4 and actual posterior means for Group 1 for dataset 1.



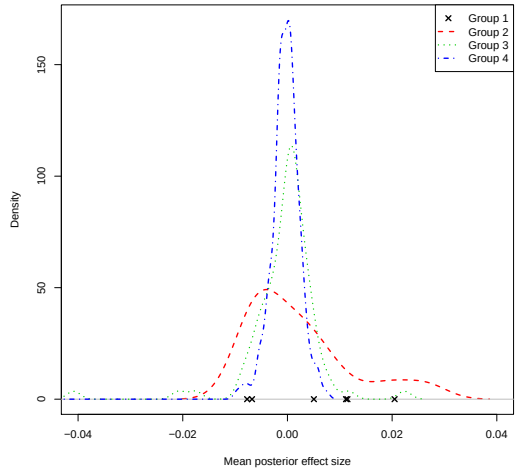
(b) Shrinkage factors for the 4 groups using dataset 1.



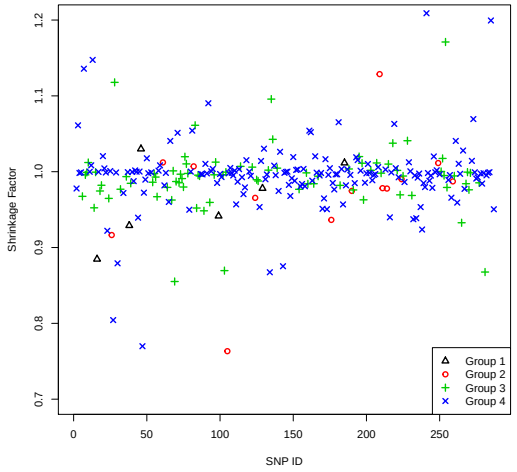
(c) Posterior mean densities for groups 2-4 and actual posterior means for Group 1 for dataset 3.



(d) Shrinkage factors for the 4 groups using dataset 3.



(e) Posterior mean densities for groups 2-4 and actual posterior means for Group 1 for dataset 8.



(f) Shrinkage factors for the 4 groups using dataset 8.

Figure 7.8: Posterior mean densities (or actual values) and shrinkage factors for Scenario 4 for LD3 with 16000 cases and 16000 controls.

Dataset	1	2	3	4	5	6	7	8	9	10
SNP46										
MLE	-0.21	-0.04	-0.25	-0.04	-0.04	-0.04	0.23	0.23	0.23	-0.21
P. Mean $LD1$	-0.03	-0.02	-0.04	-0.02	-0.02	-0.02	-0.02	-0.02	-0.02	-0.03
P. Mean $LD2$	-0.04	-0.02	-0.04	-0.02	-0.02	-0.02	-0.01	-0.01	-0.01	-0.04
P. Mean $LD3$	-0.02	-0.01	-0.02	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.02
SF P. Mean $LD1$	0.84	0.56	0.83	0.56	0.55	0.55	1.1	1.1	1.1	0.84
SF P. Mean $LD2$	0.82	0.57	0.85	0.56	0.55	0.55	1.07	1.06	1.07	0.82
SF P. Mean $LD3$	0.9	0.72	0.94	0.73	0.72	0.72	1.03	1.03	1.03	0.9
SNP47										
MLE	0.02	0.10	-0.01	0.10	0.10	0.10	-0.23	-0.23	-0.23	0.02
P. Mean $LD1$	0	-0.01	-0.01	-0.01	-0.01	-0.01	-0.02	-0.02	-0.02	0
P. Mean $LD2$	0	-0.01	-0.03	-0.01	-0.01	-0.01	-0.05	-0.05	-0.05	0
P. Mean $LD3$	0	-0.01	-0.03	-0.01	-0.01	-0.01	-0.05	-0.05	-0.05	0
SF P. Mean $LD1$	0.98	1.07	-1.24	1.07	1.07	1.07	0.9	0.91	0.9	0.98
SF P. Mean $LD2$	0.91	1.1	-3.47	1.1	1.1	1.1	0.78	0.78	0.78	0.91
SF P. Mean $LD3$	1.07	1.11	-4.65	1.11	1.11	1.11	0.76	0.77	0.77	1.06

Table 7.6: MLE, posterior mean (P.Mean), and shrinkage factors (SF) for SNP 46 and 47 applied to three LD scenarios $(LD1, LD2,$ *and* $LD3)$ *with 10 datasets having sample size of 16000 cases and 16000 controls. SNP46 is the common causal SNP. SNP47 is in very strong LD with SNP46.*

Chapter 8

The effect of incorporating FS scores into the prior for effect size in the iCOGs data

In Chapter 5 we discussed the results obtained from applying the methods (SLR, HL, Pi-MASS, and the NG prior with expected effect size variance of 1) on the iCOGs data (see Section 5.1). Here we will apply the standard NG (with expected effect size variance of 0.01) and the Modified NG ($M_1 = 0.01$ and $M_2 = 0.001$) to the iGOGs data and compare the different priors in terms of their selection of SNPs.

8.1 Results and discussion

In this section we will discuss the results obtained from applying the standard NG and the modified NG on the iCOGs data (see Section 5.1). The credible interval of the posterior distribution of effect sizes will be considered at two percentages: 85%, and 90%.

The standard NG prior was applied to the iCOGs data using MCMC with 20000 iterations, a burn-in of 2000 and thinned by 20 so the final thinned chain is 1000 observations long. Also, the modified NG prior was applied to the iCOGs data using MCMC with 35000 iterations, a burn-in of 2000 and thinned by 35 so the final thinned chain is 1000 observations long. The MCMC run for the iCOGs data took 4 days for the standard NG and a week for the modified NG on Iceberg (the University computer cluster).

Tables 8.1, presents the selected SNPs with an 85% and a 90% NG credible interval using

the standard NG prior and the modified NG prior. As expected the number of selected SNPs with the standard NG prior (the modified NG prior) decreases as the credible interval increases (6(3) selected SNPs with 85% CI, and only 3(2) selected SNPs with 90% CI). It can be seen that the standard NG selected all SNPs that were selected by the modified NG. Moreover, the three SNPs selected using the modified NG were in two different groups (Group 2, and Group 4). The modified NG does not select any SNPs at 95% and 99% sizes, whereas the standard NG selects two SNPs at 95% (SNP342, and SNP765) and a single SNP at 99% (SNP765).

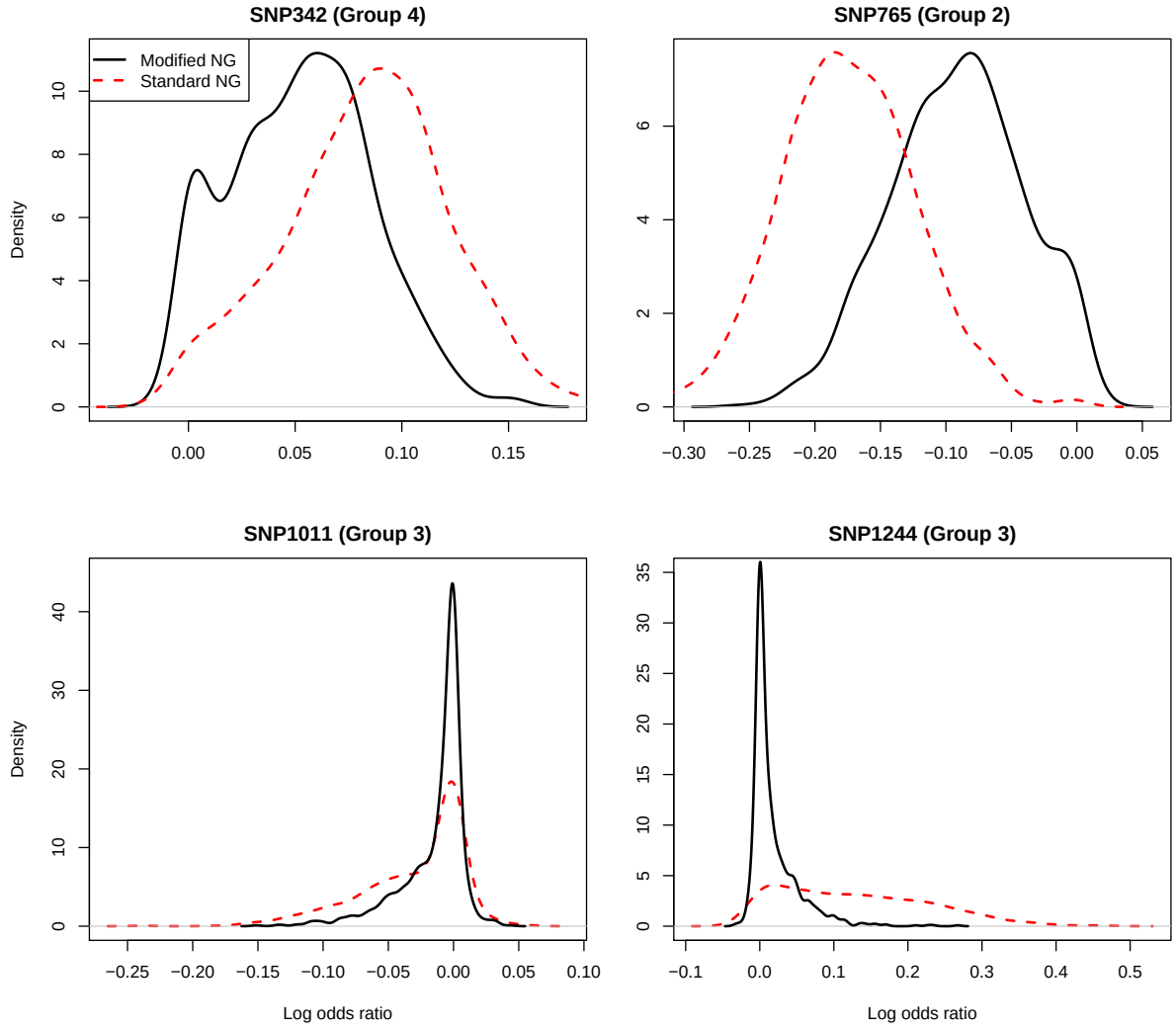


Figure 8.1: Density plots of posterior effect sizes from the standard Normal-Gamma prior and the modified Normal-Gamma prior for four SNPs. All SNPs were selected by the NG using an 85% CI. SNP342 and SNP765 were selected by the modified NG using an 85% CI, SNP1011 and SNP1244 were not when applied to the iCOGs data with 46450 cases and 42500 controls with 1733 SNPs.

Figure 8.1 shows the posterior densities of the effect sizes for four selected SNPs using a 85% CI (from Table 8.1) . Both SNP342, and SNP765 are selected by both the standard NG and the modified NG These two SNPs are located in Group 4 and Group 2 respectively. SNP1011, and SNP1244 are selected by the standard NG prior only.

It can be seen that the posterior densities of SNP342 (chosen by both priors) using the standard NG and the modified NG (SNP342) both have a lot of mass in the positive tail and very little mass around zero. For SNP765 (chosen by both NG priors) there is a lot of mass in the negative tail and there is very little mass around zero. However, for SNP1011 and SNP1244 (chosen by only the standard NG), it can be seen that the posterior densities of the modified NG is placing much more mass around zero than the standard NG.

Moreover, Figure 8.1 shows the modified NG shrinks the effect sizes much more than the standard NG. Also it can be noticed that the posterior densities of all 4 SNPs using the modified NG show varying amounts of shrinkage that depends on the SNP's group. It can be seen that there is not a SNP selected by only the modified NG. We expect that a SNP might select by the modified NG but not by the standard NG if this SNP does not shrinkage as much as the standard NG did and this might happen when the SNP located in Group 1 or Group 2. It is also worth noting that none of SNPs selected using an 85% or 90% credible interval are in FS score Group 1, whilst only one (SNP765) is in FS score group Group 2. This may be partly explained by the large sample size in the iCOGs data of approximately $90,000$.

SNPs	MAF	FS scores	SNP Group	NG	NGFS
85% CI					
SNP342	0.46	NA	4	1	1
SNP589	0.20	NA	4	1	1
SNP765	0.09	0.5	2	1	1
SNP1101	0.33	0.101	3	1	0
SNP1177	0.32	NA	4	1	0
SNP1244	0.17	0.101	3	1	0
90% CI					
SNP342	0.46	NA	4	1	1
SNP589	0.20	NA	4	1	0
SNP765	0.09	0.5	2	1	1

Table 8.1: SNPs selected using an 85% and 90% credible interval of the posterior effect sizes in the standard NG prior and the modified NG prior. We used 1 and 0 for NG, and NGFS to indicate whether the SNP in a particular method was selected or not. It is applied to the iCOGs data with 1733 SNPs and a total sample size of 89050.

Chapter 9

Discussion

The common approach to fine-mapping is to use a univariate approach. However, this approach does not take in to account the relationship between the causal variants. Therefore, here we focus on a multivariate approaches in fine-mapping studies.

In this project we not only perform a careful comparison of multivariate methods in fine-mapping, but also we incorporate functional genomic information into the continuous prior using a fully Bayesian approach in scenarios with highly correlated SNPs. The specific challenge here is the very high correlated (LD) between SNPs. A Bayesian approach offers a natural way of dealing with the high levels of multi-collinearity.

We compared the performance of both a frequentist approach (Sequential Logistic Regression (SLR)) and three Bayesian approaches: HyperLasso (HL), PiMASS and the NG prior in fine-mapping case-control studies for eight scenarios using ROC curves. It seems that there is not an outstanding approach because the performance of the methods varies in each scenario and because of the variation across the datasets.

We then developed the NG prior by including functional genomic information. The published functional significance (FS) scores were used to prioritise the causal SNP by dividing the FS scores into four groups each of which had a unique prior. These priors were chosen to control the shrinkage factors for each SNP and enable us to improve identification of the causal SNP especially when the causal SNP is placed in a high LD block.

Even when causal SNPs are placed in lower FS score groups the performance is still good. Therefore, there seems little to lose in grouping SNPs although whether performance reduces

with too many groups is unclear.

In the NG structure, the ROC results obtained by modifying the hyperparameter λ by changing the rate parameter κ are rather similar (results not shown). Thus, incorporating functional genomic information into the NG prior through allowing κ to be some function of the FS score is not a good choice. Therefore, we incorporate the functional genomic information through the expectations of the prior variance of the effect size and carefully considered suitable choices of prior variance.

In this project we considered eight scenarios. However, the results can be quite sensitive to the minor allele frequency (MAF), sample size, and odd ratios (OR). We suggest researchers consider several multivariate approaches, particularly ones that do not make hard choices. If functional information takes the form of a single score we have provided an efficient and effective way of including this in the analysis.

9.1 Limitations

One of the significant disadvantages of all the approaches is that there is a high variation across the 10 simulated datasets. This might be the reason behind unclear conclusions when we compared the four methods. Moreover, combining results to create ROC curves might not be a reasonable approach because of the variation across datasets. In addition, we used different approaches for plotting ROC curves for different methods: a "Combining Results" approach for the Bayesian approaches and "Threshold Averaging" for the frequentist approach. It might be more plausible if "Threshold Averaging" or " Vertical Averaging" were used for all methods.

Although the functional information was not complete for all SNPs, we developed an approach to incorporate the functional significant (FS) scores into the prior efficiently. However, these scores represent global scores and they do not relate to a particular disease. Therefore, this might affect the selection of causal SNPs in disease-specific path ways by placing them in low FS score groups with higher shrinkage when in fact they are likely, a priori, to be deleterious in the specific disease considered. There is no rule of thumb to select the two different expectations of the prior variance of the effect sizes (M_1, M_2). Therefore, they are chosen

semi-subjectively based on shrinkage in the breast cancer top hits data. In diseases with fewer established causal variants this becomes an even more subjective choice.

Inclusion of the FS scores into the NG prior achieved better true positive rates compared to the three Bayesian approaches without using functional information. However, the modified NG prior approach takes a lot of time to run compared to the other approaches.

In the NG prior, the expectation of the prior variance of the effect sizes (M) needs to be specified but it is not clear how to specify it. Moreover, as M decreases, smaller CI sizes are required to capture causal SNPs. As a result this choice is critical.

9.2 Future studies

In future studies, it would be interesting to compare the performance of all the methods using the ROC curves with the same method, either "Threshold Averaging" or " Vertical Averaging" rather than combining the results from different datasets. Moreover, the ENCODE database (ENCODE, 2011) provides disease-specific detailed information for each SNP although it also has a lot of missing data. Therefore, possible future research is to compare the performance by allowing inclusion of ENCODE database information in the NG prior. In addition, FS scores could be incorporated within different priors. For example it could be incorporated into PiMASS mixture priors. Another possible future study of research is to reduce the computational using the modified NG prior. To do this, we could write the code in C^{++} or integrate R and C^{++} using the $Rcpp$ package. Moreover, it is possible to attempt other shrinkage priors for our datasets, some of which are computationally simpler although offer flexibility in the shrinkage induced.

References

Abramowitz, M. and Stegun, I. A. (1964). *Handbook of mathematical functions: with formulas, graphs, and mathematical tables*, volume 55. Courier Corporation.

Ahmed, S., Thomas, G., Ghoussaini, M., Healey, C. S., Humphreys, M. K., Platte, R., Morrison, J., Maranian, M., Pooley, K. A., Luben, R., et al. (2009). Newly discovered breast cancer susceptibility loci on 3p24 and 17q23. 2. *Nature genetics*, 41(5):585–590.

Albert, J. H. and Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American statistical Association*, 88(422):669–679.

Altshuler, D. L., Durbin, R. M., Abecasis, G. R., Bentley, D. R., Chakravarti, A., Clark, A. G., Collins, F. S., la Vega, F. M. D., Donnelly, P., and Egholm, M. (2010). A map of human genome variation from population-scale sequencing. *Nature*, 467(7319):1061–1073.

Andrews, D. F. and Mallows, C. L. (1974). Scale mixtures of normal distributions. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 99–102.

Balding, D. J. (2006). A tutorial on statistical methods for population association studies. *Nature Reviews Genetics*, 7(10):781–791.

Bazaraa, M. S., Sherali, H. D., and Shetty, C. M. (2013). *Nonlinear programming: theory and algorithms*. John Wiley & Sons.

Bhattacharya, A., Pati, D., Pillai, N. S., and Dunson, D. B. (2012). Bayesian shrinkage. *arXiv preprint arXiv:1212.6088*.

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

Boggis, E., Milo, M., and Walters, K. (2016). eQuIPS: eQTL analysis using informed partitioning of SNPs—-a fully bayesian approach. *Genetic epidemiology*, 40(4):273–283.

Breymann, W. and Lüthi, D. (2013). ghyp: A package on generalized hyperbolic distributions. *Available in http://cran. r-project. org/web/packages/ghyp/vignettes/Generalized_Hyperbolic_Distribution. pdf*.

Cai, Q., Zhang, B., Sung, H., Low, S.-K., Kweon, S.-S., Lu, W., Shi, J., Long, J., Wen, W., Choi, J.-Y., et al. (2014). Genome-wide association analysis in east asians identifies breast cancer susceptibility loci at 1q32. 1, 5q14. 3 and 15q26. 1. *Nature genetics*, 46(8):886–890.

Casella, G. and Berger, R. L. (2002). *Statistical inference*, volume 2. Duxbury, Pacific Grove, CA.

Devlin, B. and Risch, N. (1995). A comparison of linkage disequilibrium measures for fine-scale mapping. *Genomics*, 29(2):311–322.

Easton, D. F., Pooley, K. A., Dunning, A. M., Pharoah, P. D., Thompson, D., Ballinger, D. G., Struewing, J. P., Morrison, J., Field, H., Luben, R., et al. (2007). Genome-wide association study identifies novel breast cancer susceptibility loci. *Nature*, 447(7148):1087–1093.

Embrechts, P. (1983). A property of the generalized inverse gaussian distribution with some applications. *Journal of Applied Probability*, pages 537–544.

ENCODE (2011). A user's guide to the encyclopedia of DNA elements (ENCODE). *PLoS biology*, 9(4):e1001046.

Fachal, L. and Dunning, A. M. (2015). From candidate gene studies to GWAS and post–GWAS analyses in breast cancer. *Current opinion in genetics & development*, 30:32–41.

Fawcett, T. (2006a). An introduction to ROC analysis. *Pattern recognition letters*, 27(8):861–874.

Fawcett, T. (2006b). An introduction to ROC analysis. *Pattern recognition letters*, 27(8):861–874.

Fletcher, O., Johnson, N., Orr, N., Hosking, F. J., Gibson, L. J., Walker, K., Zelenika, D., Gut, I., Heath, S., Palles, C., et al. (2011). Novel breast cancer susceptibility locus at 9q31. 2: results of a genome-wide association study. *Journal of the National Cancer Institute*, 103(5):425–435.

Foulkes, A. S. (2009). *Applied statistical genetics with R: for population-based association studies*. Springer Science & Business Media.

French, J. D., Ghoussaini, M., Edwards, S. L., Meyer, K. B., Michailidou, K., Ahmed, S., Khan, S., Maranian, M. J., O'Reilly, M., Hillman, K. M., et al. (2013). Functional variants at the 11q13 risk locus for breast cancer regulate cyclin d1 expression through long-range enhancers. *The American Journal of Human Genetics*, 92(4):489–503.

Fridley, B. L., Iversen, E., Tsai, Y.-Y., Jenkins, G. D., Goode, E. L., and Sellers, T. A. (2011). A latent model for prioritization of SNPs for functional studies. *PloS one*, 6(6):e20764.

Gelman, A. and Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, pages 457–472.

Gentle, J. E. (2009). *Computational Statistics*, volume 308. Springer.

Ghoussaini, M., Fletcher, O., Michailidou, K., Turnbull, C., Schmidt, M. K., Dicks, E., Dennis, J., Wang, Q., Humphreys, M. K., Luccarini, C., et al. (2012). Genome-wide association analysis identifies three new breast cancer susceptibility loci. *Nature genetics*, 44(3):312–318.

Glubb, D. M., Maranian, M. J., Michailidou, K., Pooley, K. A., Meyer, K. B., Kar, S., Carlebur, S., O'Reilly, M., Betts, J. A., Hillman, K. M., et al. (2015). Fine-scale mapping of the 5q11. 2 breast cancer locus reveals at least three independent risk variants regulating map3k1. *The American Journal of Human Genetics*, 96(1):5–20.

Griffin, J. E. and Brown, P. J. (2010). Inference with normal-gamma prior distributions in regression problems. *Bayesian Analysis*, 5(1):171–188.

Guan, Y. and Stephens, M. (2011). Bayesian variable selection regression for genome-wide association studies and other large-scale problems. *The Annals of Applied Statistics*, 5(3):1780–1815.

Hastings, W. K. (1970). Monte carlo sampling methods using markov chains and their applications. *Biometrika*, 57(1):97–109.

Hill, W. and Weir, B. (1994). Maximum-likelihood estimation of gene location by linkage disequilibrium. *American journal of human genetics*, 54(4):705.

Hoggart, C. J., Whittaker, J. C., De Iorio, M., and Balding, D. J. (2008). Simultaneous analysis of all SNPs in genome-wide and re-sequencing association studies. *PLoS genetics*, 4(7):e1000130.

Kichaev, G., Yang, W.-Y., Lindstrom, S., Hormozdiari, F., Eskin, E., Price, A. L., Kraft, P., and Pasaniuc, B. (2014). Integrating functional data to prioritize causal variants in statistical fine-mapping studies. *PLoS genetics*, 10(10):e1004722.

Lee, P. H. and Shatkay, H. (2009). An integrative scoring system for ranking SNPs by their potential deleterious effects. *Bioinformatics*, 25(8):1048–1055.

Lewontin, R. (1964). The interaction of selection and linkage. I. general considerations; heterotic models. *Genetics*, 49(1):49.

Marchini, J. and Howie, B. (2010). Genotype imputation for genome-wide association studies. *Nature Reviews Genetics*, 11(7):499–511.

Mengersen, K. L., Robert, C. P., and Guihenneuc-Jouyaux, C. (1999). Mcmc convergence diagnostics: a reviewww. *Bayesian statistics*, 6:415–440.

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

Michailidou, K., Beesley, J., Lindstrom, S., Canisius, S., Dennis, J., Lush, M. J., Maranian, M. J., Bolla, M. K., Wang, Q., Shah, M., et al. (2015). Genome-wide association analysis of more than 120,000 individuals identifies 15 new susceptibility loci for breast cancer. *Nature genetics*, 47(4):373–380.

Michailidou, K., Hall, P., Gonzalez-Neira, A., Ghoussaini, M., Dennis, J., Milne, R. L., Schmidt, M. K., Chang-Claude, J., Bojesen, S. E., Bolla, M. K., et al. (2013a). Large-scale genotyping identifies 41 new loci associated with breast cancer risk. *Nature genetics*, 45(4):353–361.

Michailidou, K., Hall, P., Gonzalez-Neira, A., Ghoussaini, M., Dennis, J., Milne, R. L., Schmidt, M. K., Chang-Claude, J., Bojesen, S. E., Bolla, M. K., et al. (2013b). Large-scale genotyping identifies 41 new loci associated with breast cancer risk. *Nature genetics*, 45(4):353–361.

Miller, J. J. (1977). Asymptotic properties of maximum likelihood estimates in the mixed model of the analysis of variance. *The Annals of Statistics*, pages 746–762.

Milne, R. L., Burwinkel, B., Michailidou, K., Arias-Perez, J.-I., Zamora, M. P., Menéndez-Rodríguez, P., Hardisson, D., Mendiola, M., González-Neira, A., Pita, G., et al. (2014). Common non-synonymous SNPs associated with breast cancer susceptibility: findings from the breast cancer association consortium. *Human molecular genetics*, 23(22):6096–6111.

Nelder, J. A. and Baker, R. (1972). Generalized linear models. *Encyclopedia of Statistical Sciences*.

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

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., Müller, M., and Robin, M. X. (2014). Package 'pROC'.

Saccone, S. F., Saccone, N. L., Swan, G. E., Madden, P. A., Goate, A. M., Rice, J. P., and

Bierut, L. J. (2008). Systematic biological prioritization after a genome-wide association study: an application to nicotine dependence. *Bioinformatics*, 24(16):1805–1811.

Siddiq, A., Couch, F. J., Chen, G. K., Lindström, S., Eccles, D., Millikan, R. C., Michailidou, K., Stram, D. O., Beckmann, L., Rhie, S. K., et al. (2012). A meta-analysis of genome-wide association studies of breast cancer identifies two novel susceptibility loci at 6q14 and 20q11. *Human molecular genetics*, 21(24):5373–5384.

Sing, T., Sander, O., Beerenwinkel, N., and Lengauer, T. (2005). Rocr: visualizing classifier performance in r. *Bioinformatics*, 21(20):3940.

Spencer, A. V., Cox, A., Lin, W.-Y., Easton, D. F., Michailidou, K., and Walters, K. (2016). Incorporating functional genomic information in genetic association studies using an empirical bayes approach. *Genetic epidemiology*, 40(3):176–187.

Stacey, S. N., Manolescu, A., Sulem, P., Rafnar, T., Gudmundsson, J., Gudjonsson, S. A., Masson, G., Jakobsdottir, M., Thorlacius, S., Helgason, A., et al. (2007). Common variants on chromosomes 2q35 and 16q12 confer susceptibility to estrogen receptor-positive breast cancer. *Nature genetics*, 39(7):865.

Stacey, S. N., Manolescu, A., Sulem, P., Thorlacius, S., Gudjonsson, S. A., Jonsson, G. F., Jakobsdottir, M., Bergthorsson, J. T., Gudmundsson, J., Aben, K. K., et al. (2008). Common variants on chromosome 5p12 confer susceptibility to estrogen receptor–positive breast cancer. *Nature genetics*, 40(6):703–706.

Strachan, T. and Read, A. P. (2004). Human molecular genetics. *Garland Science, New York*, 635.

Su, Z., Marchini, J., and Donnelly, P. (2011). HAPGEN2: simulation of multiple disease SNPs. *Bioinformatics*, 27(16):2304–2305.

Szklo, M. and Nieto, F. J. (2012). *Epidemiology: beyond the basics*. Jones & Bartlett Publishers.

Thomas, G., Jacobs, K. B., Kraft, P., Yeager, M., Wacholder, S., Cox, D. G., Hankinson, S. E., Hutchinson, A., Wang, Z., Yu, K., et al. (2009). A multistage genome-wide association study in breast cancer identifies two new risk alleles at 1p11. 2 and 14q24. 1 (rad51l1). *Nature genetics*, 41(5):579–584.

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

Turnbull, C., Ahmed, S., Morrison, J., Pernet, D., Renwick, A., Maranian, M., Seal, S., Ghoussaini, M., Hines, S., Healey, C. S., et al. (2010). Genome-wide association study identifies five new breast cancer susceptibility loci. *Nature genetics*, 42(6):504–507.

Udler, M. S., Ahmed, S., Healey, C. S., Meyer, K., Struewing, J., Maranian, M., Kwon, E. M., Zhang, J., Tyrer, J., Karlins, E., et al. (2010). Fine scale mapping of the breast cancer 16q12 locus. *Human molecular genetics*, page ddq122.

VanLiere, J. M. and Rosenberg, N. A. (2008). Mathematical properties of the r2 measure of linkage disequilibrium. *Theoretical population biology*, 74(1):130–137.

Wakefield, J. (2008). Reporting and interpretation in genome-wide association studies. *International journal of epidemiology*, 37(3):641–653.

Wakefield, J. (2009). Bayes factors for genome-wide association studies: comparison with p-values. *Genetic epidemiology*, 33(1):79–86.

West, M. (1987). On scale mixtures of normal distributions. *Biometrika*, pages 646–648.

Zheng, W., Long, J., Gao, Y.-T., Li, C., Zheng, Y., Xiang, Y.-B., Wen, W., Levy, S., Deming, S. L., Haines, J. L., et al. (2009). Genome-wide association study identifies a new breast cancer susceptibility locus at 6q25. 1. *Nature genetics*, 41(3):324–328.

Appendices

Appendix A

Trace and ACF plots

Here we will show the trace and ACF plots for the posterior of $\beta_{589}, \psi_{589}, \lambda$ and γ^2 using Normal-Gamma method for asymptotic normal likelihood. Applying iCOGs data with 46450 cases and 42500 controls with 1733 SNPs that was mentioned in Section 5.3.

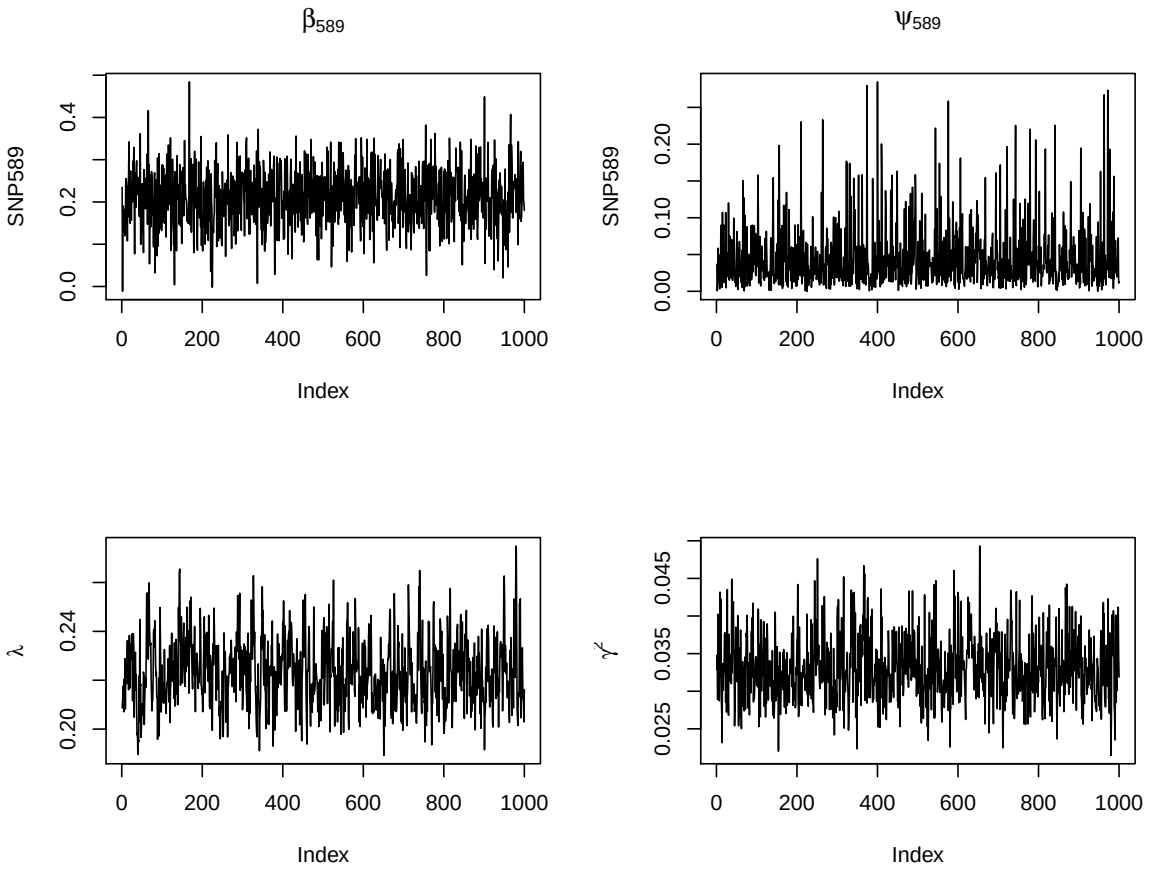


Figure A.1: Trace plots for the posterior of $\beta_{589}, \psi_{589}, \lambda$ and γ^2 using Normal-Gamma method for asymptotic normal likelihood. Applying iCOGs data with 46450 cases and 42500 controls with 1733 SNPs. In Normal-Gamma method for asymptotic normal likelihood, the MCMC run for 20,000 iterations with 2,000 burn-in and thinning by 20.

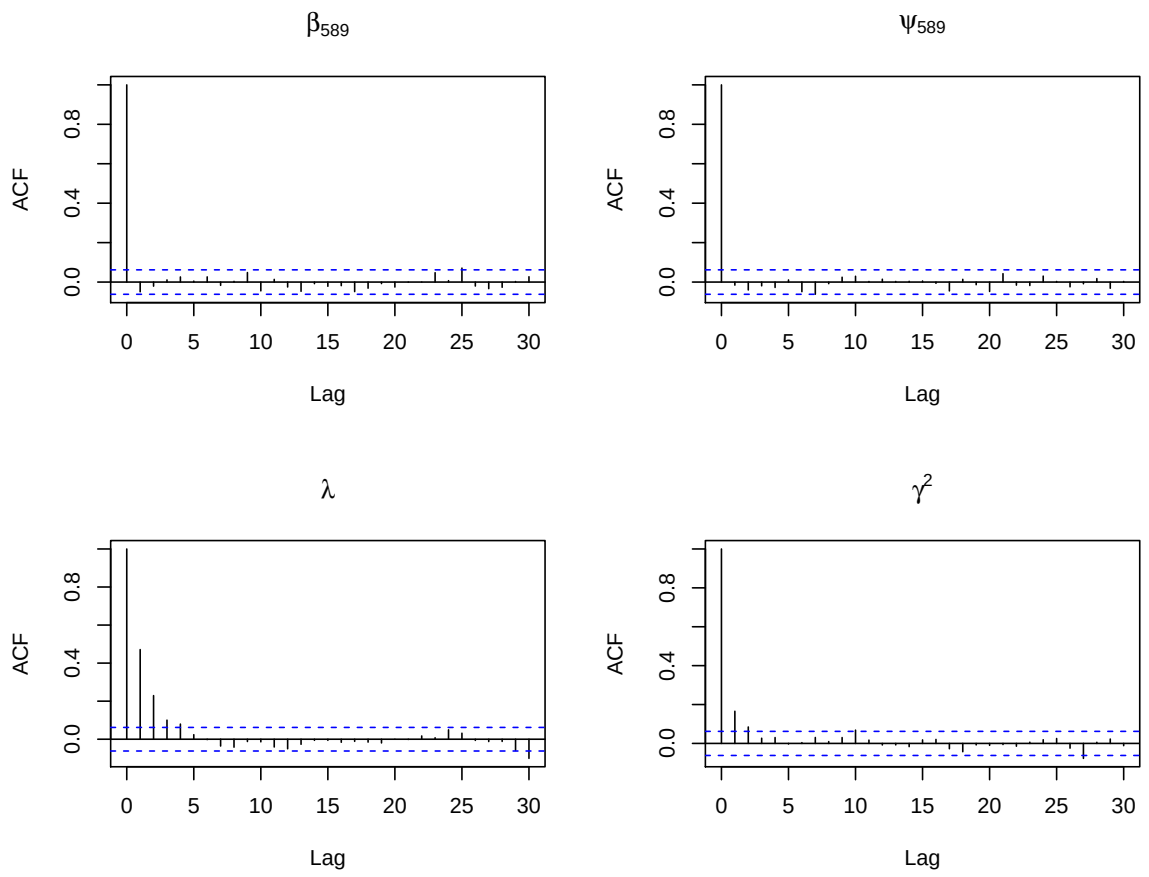


Figure A.2: ACF plots for the posterior of $\beta_{589}, \psi_{589}, \lambda$ and γ^2 using Normal-Gamma method for asymptotic normal likelihood. Applying iCOGs data with 46450 cases and 42500 controls with 1733 SNPs. In Normal-Gamma method for asymptotic normal likelihood, the MCMC run for 20,000 iterations with 2,000 burn-in and thinning by 20.