Author/s:
Xu, Zemei

Title:
Sparse Bayesian Grouped Regression with Application to Genome-Wide Association Studies

Date:
2019

Persistent Link:
https://hdl.handle.net/11343/233804

Sparse Bayesian Grouped Regression with Application to Genome-Wide Association Studies

Thesis by

Zemei Xu

Submitted in Partial Fulfillment of the Requirements for the
Degree of
Doctor of Philosophy

Melbourne School of Population and Global Health
Melbourne School of Mathematics and Statistics
The University of Melbourne

2nd March 2019

Abstract

Statistical variable selection, also known as feature selection, has become an indispensable tool in many research areas involving machine learning and data mining. The object of statistical variable selection is to select the best subset of predictors for fitting or predicting the response variable from a potentially large collection of candidate predictors. It is particularly important in high-dimensional problems such as cancer genetics, where there are potentially thousands of predictors and only a few are associated with the outcome. Moreover, predictors may have underlying group structures in them, so it is desirable to take the underlying group effects into consideration when performing variable selection and handle the grouping structure present in the model when selecting important variables.

This thesis presents a Bayesian grouped regression model with continuous global-local shrinkage priors to tackle the high-dimensional variable selection problem by introducing shrinkage parameters at the group level as well as within each group. This model is able to handle complex group hierarchies that include overlapping and multilevel group structures, and also enjoys the advantages of handling sparsity as well as strong signals. The proposed method is also applied to a real high-dimensional GWAS breast cancer dataset, Haiman-Hopper dataset, to analyse the breast cancer genome-wide association study.

Thesis Declaration

This is to certify that:

i. The thesis comprises only my original work towards Doctor of Philosophy except where indicated in the Preface;

ii. Due acknowledgement has been made in the text to all other materials used;

iii. The thesis is fewer than 100 000 words in length, exclusive of tables, maps, bibliographies and appendices.

Zemei Xu

March, 2019

Preface

This thesis presents the work that I have completed throughout my PhD candidature at both Centre for Epidemiology and Biostatistics, Melbourne School of Population and Global Health, and Melbourne School of Mathematics and Statistics, University of Melbourne, Australia.

The work presents in Chapter 4 was accomplished in collaboration with Dr. Daniel F. Schmidt, Dr. Enes Makalic, Associate Prof. Guoqi Qian and Prof. John L. Hopper. I presented two extensions of the Bayesian horseshoe model, set up the simulation studies and applied to the real datasets. Dr. Daniel F. Schmidt and Dr. Enes Makalic provided guidance and a lot of useful suggestions on simulation studies and real data implementations. Associate Prof. Guoqi Qian and Prof. John L. Hopper greatly assisted me in reviewing the publication and provided advice.

Part of this work has been published at:

Z. Xu, D. F. Schmidt, E. Makalic, G. Qian, and J. L. Hopper, "Bayesian grouped horseshoe regression with application to additive models," in *Australasian Joint Conference on Artificial Intelligence*, pp. 229–240, Springer, 2016.

For Chapter 5, my original contribution includes development of grouped Bayesian model for complex group structures, extension of decouple shrinkage and selection method to grouped variables, proposal of generalised information criteria with an alternative degrees of freedom estimate, set-up of simulation tests and application to real data. Dr. Daniel F. Schmidt and Dr. Enes Makalic contributed part of the ideas and provided substantially useful discussions.

The work presents in Chapter 6 was accomplished in collaboration with Dr. Daniel F. Schmidt, Dr. Enes Makalic, Associate Prof. Guoqi Qian and Prof. John L. Hopper. I proposed to apply posterior likelihood ratio to Byesian grouped model, set up the model, conduct the simulation and applied to high-dimensional real data through the supercomputing system. Dr. Daniel F. Schmidt and Dr. Enes

Makalic contributed part of the ideas and assisted in the design of the simulation tests. Associate Prof. Guoqi Qian involved in some useful discussions. Prof. John L. Hopper provided the high-dimensional breast cancer dataset.

Acknowledgements

First of all, I would like to gratefully and sincerely thank my supervisors Dr. Daniel F. Schmidt, Dr. Enes Makalic, A/Prof. Guoqi Qian and Prof. John L. Hopper for their support, guidance, patience and advice throughout my PhD. Their guidance helped me in all the time of research and writing of this thesis. Special thanks to Daniel and Enes for their endless support and care about my work, and always being respondent to my questions promptly and patiently. Indeed, without their guidance, I would not be able to put the topic together.

I would also like to thank my dear friends and colleagues from both Centre for Epidemiology and Biostatistics and Melbourne School of Mathematics and Statistics for being supportive and encouraging all the time.

Last but not least, I would like to thank my families for their unconditional support, both financially and emotionally throughout my degree. In particular, the patience and understanding shown by them are greatly appreciated.

Thank you all for your unwavering support.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Introduction

Regression is a standard tool for examining the relationship between a set of independent variables and a response variable. It is widely used as a tool to estimate parameters in a statistical model and to make predictions. In addition to model estimation and prediction, parsimony is another important aspect for a regression model as many variables may be redundant and only associated predictors are of interest.

Variable selection has appeared in many research fields such as machine learning and data mining. The aim of statistical variable selection is to select the best subset of predictors for fitting or predicting the response variable from a potentially large collection of candidate predictors. It is particularly important in high-dimensional problems such as cancer genetics, where there are millions of predictors and only a few are associated with the outcome. There are many existing methods for variable selection, such as subset selection [2, 3] which includes forward selection [4], backward selection [5], forward stagewise selection [6], and

penalised regression which includes non-negative garrote [7], ridge [8], least absolute shrinkage and selection operator (Lasso) [9], etc. The penalised regression is an alternative to a standard linear regression and it minimises a loss function plus an extra penalty term. The penalty term involves a regularisation parameter which controls the size of the coefficients and the amount of the shrinkage.

There have often been concerns that some predictors may have underlying group structures. When performing variable selection by selecting one predictor at a time, we may miss the potential group effects in the predictors, and therefore, techniques that are able to perform group selection are preferable. For example, gene data contains biological information; single nucleotide polymorphisms (SNPs) can be grouped into genes and genes can be grouped into pathways. Instead of selecting individual SNPs, sometimes researchers are more interested in selecting genes which implies selecting a group of SNPs. Things become more complicated when groups are overlapped which is common in real world applications. For example, in statistical genomics, predictors corresponding to single nucleotide polymorphisms (SNPs) may belong to different genes, and a collection of genes may form a biological pathway. Additionally, pathways may share genes and genes may share single nucleotide polymorphisms. By utilising an appropriate number of grouping levels, the complexity of this situation is easily handled by our proposed hierarchical model. So each predictor may belong to more than one group and the groups are overlapped. In this scenario, a model that is able to handle multilevel and overlapping group structures is desired.

There are a few group variable selection techniques. For example, non-Bayesian penalised regression such as group non-negative garrotte [7], group Lasso [10], adaptive group Lasso [11]. However non-Bayesian penalised regression involves estimation of regularisation parameter and it does not naturally provide confidence intervals for parameter estimation.

An important alternative class of penalised regression method is Bayesian penalised regression. Bayesian methods automatically provide estimation for regularisation parameters and produces samples of coefficients that can be further used

to calculate confidence intervals. There are two general types of priors for Bayesian regression: spike and slab priors and continuous priors. The spike and slab priors heavily rely on Markov Chain Monte Carlo (MCMC) algorithms and become computationally inefficient especially for high-dimensional data. The continuous shrinkage priors can be expressed as Gaussian scale mixtures [12, 13], which is a class that contains many well-known distributions such as the t distribution, the double-exponential distribution, the normal-exponential-gamma distribution, and the horseshoe distribution. Thus, lead to a simpler MCMC algorithm. They enable simultaneous variable selection and estimation. Bayesian regression with continuous shrinkage priors are our main focus. There are many examples: Bayesian ridge, Bayesian Lasso, Bayesian horseshoe, and Bayesian horseshoe+.

One potential issue with Bayesian regression with continuous priors is that it does not automatically yield a sparse solution. A sparse solution refers to the model that uses only a few variables, while most coefficients of variables have no impact on it and are set to zero. Thus, the Bayesian regression with continuous priors requires a further step to generate sparse estimates. Some of the most common procedures of sparsification are simple thresholding rules, penalised variable selection based on posterior credible regions, and decoupled shrinkage and selection method (DSS) [14]. The DSS finds the best sparse estimates through a trade-off between the number of variables in the model and the predictive performance in terms of the mean squared prediction error with correlation of predictors being taken into consideration. The DSS produces a solution path of sparse estimates, and then some criteria are applied to the solution path to select the final sparse model. However, the DSS has not been extended to handle grouped variables.

This thesis presents a Bayesian grouped regression model with continuous shrinkage priors that handles multilevel and overlapping group structures of underlying data by introducing shrinkage parameters at the group level as well as within each group. After obtaining posterior samples from the Bayesian model, we propose the group decoupled shrinkage and selection method that applies to grouped variables and produces a sparse solution path using the group non-negative garrotte.

We then propose the generalised information criteria for group DSS with an alternative degrees of freedom to select the final sparsified model. This algorithm enjoys the advantages of handling sparsity as well as strong signals by leaving strong signals unshrunk and penalising noise variables severely, produces sparse estimation with good predictive performance, and handles the high-dimensional group variable selection problem.

The proposed method is then applied to a genome-wide association study (GWAS) data on breast cancer. GWAS is one of the most common methods of exploring the genetic basis of phenotypic traits. In general, GWAS is a study of people's DNA to determine what mutations people have that might cause an increase in the risk of getting a certain disease. It works by scanning the entire genome to attempt to find out what genetic variants might be associated with a particular disease. Those genetic variants are usually referred to as single nucleotide polymorphisms (SNPs). The significance of these studies is that once SNPs are identified, people can then develop better approaches to identify the causative mutations, and therefore, have a deeper knowledge of developing preventative strategies for the disease. In addition, it helps scientists to invent new drugs for treating or curing a particular disease.

However, with millions of SNPs, it is difficult to build a suitable model that identifies important SNPs in the absence of a large sample size. Standard approaches to analyse data are marginal methods where either the χ^2 distribution or logistic regression is applied to the data. The conventional marginal testing looks at each single SNP one-by-one and tests the associations for each SNP individually according to some significance threshold. After performing the Bonferroni adjustment for multiple tests, the threshold is divided by the total number of tests performed to become the new stringent threshold of significance. Standard approaches have several potential limitations. First, they tend to be overly conservative. With a stringent threshold of significance, the current procedures tend to heavily penalise false discoveries and therefore might not include features which indeed have an association with the disease. Second, very often, when two SNPs are inherited

together in an offspring, they might be highly correlated and are in linkage disequilibrium (LD). Marginal methods fail to separate SNPs in LD appropriately and neglect linkage disequilibrium when performing the tests. Third, current procedures fit each SNP separately rather than fitting all SNPs together, fails to take combined multiple joint effects of SNPs into consideration, which also results in an overly simple model of association.

We analyse a breast cancer GWAS by applying our proposed Bayesian group model to a real high-dimensional breast cancer dataset, Haiman-hopper (H2) dataset. The model looks at all SNPs within a gene simultaneously, takes group effects into consideration, and picks important features that are associated with the disease based on p-values from posterior likelihood ratio (PLR) method.

1.2 Contributions

The original contributions of this thesis are outlined below:

- Proposed two extensions of the regular Bayesian horseshoe: the grouped Bayesian horseshoe and the hierarchical Bayesian horseshoe to deal with grouped variables by introducing shrinkage parameters at the group level as well as within each group (Chapter 4). The model was applied to additive models where group structures naturally exist to test the performance. The performance of proposed models applied to additive models as well as real datasets were presented.

- Proposed a generalisation of grouped Bayesian model with continuous global-local shrinkage priors, to handle complex group structures that include overlapping and multilevel group structures; in particular, we extended the horseshoe and horseshoe+ priors for complex grouped predictor structures (Chapter 5).

- Developed group decouple shrinkage and selection by extending the decoupled shrinkage and selection method to handle grouped variables (Chapter 5).

The group non-negative garrotte was incorporated in group DSS to select a number of candidate sparse solutions for grouped variables.

- Proposed the generalised information criteria for grouped DSS method (Chapter 5).
- Proposed an alternative estimate for degrees of freedom by using posterior expected estimates of the degrees of freedom of the models in the group DSS model path to select the final sparse model, and to perform group-wise selection. Incorporated extended Bayes information criteria into our proposed generalised information criteria method and conducted a series of simulation studies to illustrate the performance of group selections as well as individual selection under different scenarios (Chapter 5).
- Developed a gene-based testing method to detect important features of GWAS. It combined posterior likelihood ratio method with the our proposed Bayesian grouped model (Chapter 6).
- Applied proposed gene-based testing method to a high-dimensional real GWAS breast cancer dataset, Haiman-Hopper dataset, to perform gene-based testing througth a super-computing platform. The results of proposed method were compared with existing gene-based testing methods (Chapter 6).
- Built a predictive model based on SNPs which can be potentially used for foreseeing people with possible risk of getting breast cancer (Chapter 6).

1.3 Thesis Outline

The organization of this thesis is as follows.

Chapter 2 is devoted to the literature review of Bayesian inference, linear regression models, non-Bayesian penalised regression and Bayesian linear regression. It covers some most popular modern statistical regression techniques. In addition,

it introduces the group structures and some modern techniques for the Bayesian and non-Bayesian group models.

In Chapter 3, we cover the topics of genome-wide association studies which include the conventional approach for GWAS and recent statistical methods for GWAS. More specifically, breast cancer literature is reviewed, and current findings for breast cancer is stated.

Chapter 4 presents a grouped Bayesian horseshoe model that handles grouped variables by introducing shrinkage parameters at the group level as well as within each group, and the application of proposed methods to the important class of additive models where group structures naturally exist is also showed.

Chapter 5 presents a general Bayesian group continuous shrinkage priors that handles complex group structures. The extension of the recently proposed decoupled shrinkage and selection technique to the problem of selecting groups of variables from posterior samples is also discussed. A final, sparse model, the generalised information criteria approaches to the DSS framework with an alternative degrees of freedom estimator for sparse Bayesian linear models that takes into account the effects of shrinkage on the model coefficients is presented. Simulations and real data analysis of proposed method in terms of correct identification of active and inactive groups, and prediction, are compared with a Bayesian grouped slab-and-spike approach.

Chapter 6 presents the construction of posterior likelihood ratio method applied to our proposed Bayesian sparse global-local shrinkage regression for grouped variables selection. The proposed method can be easily applied in conjunction with sparsity construction of our Bayesian model. Simulation results of PLR is compared with a conventional likelihood ration test. The application of proposed method to a real GWAS breast cancer dataset, a Haiman-Hopper dataset, is presented. The results of our findings with existing literatures in breast cancer studies, as well as results from existing gene-based testing methods for H2 dataset are compared. A predictive model for H2 dataset is also constructed.

Chapter 7 contains a final conclusion and summary of the thesis, together with a discussion of possible future work. This chapter is followed by an appendix that contains a list of selected SNPs for Haiman-Hopper dataset, and a list of references used in this thesis.

Chapter 2

Statistical Inference

There are two main approaches in statistics: the frequentist approach and the Bayesian approach. The frequentist approach is where data are randomly and repeatedly sampled while the underlying parameters are remain fixed. It develops procedures to see the performance for all possible random samples. For Bayesian approach, data are observed so that it is fixed from the realised sample and parameters are unknown quantities that are treated as a random variable with its distribution of possible values. It offers different advantages over the widely used frequentist approach.

In this chapter, we cover the review of basic Bayesian inference, linear regression, logistic regression, penalised regression methods, Bayesian linear methods, and grouping structures.

2.1 Bayesian Inference

The complete method of Bayesian statistical inference has been developed based on Bayes' theorem. The basis of the Bayesian approach is that the true value of the parameters are unknown and are treated as random variables. To make inference about the parameters through prior knowledge and the observed data,

the rules of probability are then applied directly to gain revised belief about the parameters.

2.1.1 Bayes' Theorem

In statistics, Bayes' theorem, also known as Bayes' rule [15], is the conditional probability or a posterior probability of an event given some other event occurred. The mathematical expression of the Bayes' theorem is:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(B|A)P(A)}{P(B)}, \tag{2.1}$$

where both A and B are events and $P(\cdot)$ is a probability. In many situations, it is difficult to compute $P(A|B)$ directly, while information about $P(B|A)$ is available, and therefore, the Bayes' theorem allows for computation of $P(A|B)$ through $P(B|A)$.

A more general formula of (2.1) applies to partitions of a sample space. Let $A_1, A_2, \ldots, A_m$ be partitions of the sample space S, which means that they are mutually exclusive and exhaustive, i.e.:

$$A_j \cap A_k = \emptyset \quad \text{and} \quad A_1 \cup A_2 \cup \cdots \cup A_m = S, \quad j, k \in \{1, 2, \ldots, m\}$$

then the posterior probability of event A_i given event B has occurred becomes:

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{\sum_{j=1}^{\infty} P(B|A_j)P(A_j)}, \quad i = 1, 2, \ldots, m. \tag{2.2}$$

Bayes' theorem is central to Bayesian inference.

2.1.2 Bayesian Inference

The Bayesian inference is a statistical inference which is based on Bayes' theorem. It is used to update the posterior probability given the observed data. The basis for Bayesian inference is derived from Bayes' theorem (2.1) by replacing event A

with parameters of interest, event B with observations $\mathbf{Y}$ and probabilities with densities p.

Let θ be an unknown parameter of interest which can also be a vector, and it indexes a population where a random sample $Y_1, \ldots, Y_n$ is drawn from. Then the model-based formulation of Bayes' theorem results in the following form:

$$p(\theta|\mathbf{y}) = \frac{p(\mathbf{y}|\theta)p(\theta)}{p(\mathbf{y})}, \tag{2.3}$$

where $p(\theta)$ is the prior distribution of the parameter before the data is seen, the data distribution $p(\mathbf{y}|\theta)$ is the likelihood of $\mathbf{y}$ under a model, $p(\theta|\mathbf{y})$ is the full posterior distribution and the denominator $p(\mathbf{y})$ is the marginal likelihood of $\mathbf{y}$ which can be written as:

$$p(\mathbf{y}) = \int p(\mathbf{y}|\theta)p(\theta)d\theta \tag{2.4}$$

in the continuous case; and

$$p(\mathbf{y}) = \sum_{\theta} p(\mathbf{y}|\theta)p(\theta) \tag{2.5}$$

in the discrete case. In Bayesian statistics, Bayesian inference is unable to be conducted if $p(\theta|\mathbf{y})$ cannot be normalised into a probability density function. So if the marginal likelihood of $\mathbf{y}$ normalises the posterior distribution, it ensures that the posterior distribution is a proper distribution and integrates to one. The marginal likelihood $p(\mathbf{y})$ can then be set to an unknown constant c, and (2.3) becomes:

$$p(\theta|\mathbf{y}) = \frac{p(\mathbf{y}|\theta)p(\theta)}{c} \propto p(\mathbf{y}|\theta)p(\theta).$$

Therefore, to effectively obtain a posterior distribution, simply multiply the prior distribution by the likelihood, remove all constants from prior and likelihood that does not depend on parameter θ, and finally determine the normalising constant that forces the expression to integrate to one.

Given the prior distribution and the observed distribution which is function of parameter values, the posterior distribution $p(\theta|\mathbf{y})$ in (2.3) is derived using the

Bayes' rule, which is computed as the likelihood function of the observed distribution, multiplied by the prior distribution, and divided by a normalising constant.

2.1.2.1 Prior Distributions

Bayes' theorem provides a feasible approach to update beliefs about the parameter after the data is seen. A prior belief about the parameter must be given before the data is observed, and therefore a prior distribution must be specified. Based on prior knowledge of the problem, some prior information about θ is obtained and its variation can be described by a probability distribution, known as the *prior distribution*. If the prior distribution $p(\theta)$ is based on prior belief or past information and is given before the data is taken into account, it is subjective. If there is no prior information about the parameter, the prior distribution should minimise the impact on the inference, and thus it is objective.

The prior distribution affects the posterior distribution and therefore choosing an appropriate prior distribution becomes crucial in Bayesian inference. There are mainly two categories of prior distributions: informative priors and uninformative priors.

Informative Prior When the prior information or knowledge about parameter θ exist before data is observed, or when the opinion of statisticians or experts is available, it should be incorporated in constructing the prior distribution so that it reflects such information or beliefs about the parameter θ. Such information may come from previous experiments or past experience. The informative prior thus dominates the likelihood function of the data and has an influence on the posterior distribution. The more informative the prior, the more data is needed to change the beliefs about the parameter.

Uninformative Prior If there is no prior information available, a prior distribution with minimum impact on the inference is desired. We call such prior distribution an uninformative prior. The uninformative prior is objective and the

posterior distribution with an uninformative prior is data-driven. A general way to construct uninformative priors is to use the well-known Jeffreys prior [16]:

$$p(\theta) \propto I^{\frac{1}{2}}(\theta), \tag{2.6}$$

where $I(\theta)$ is the expected Fisher information:

$$I(\theta) = \mathrm{E}\left[\frac{\partial^2}{\partial\theta^2}\log p(\mathbf{y}|\theta)\right].$$

In the multivariate case, the Jeffreys prior is written as:

$$p(\boldsymbol{\theta}) \propto |I(\boldsymbol{\theta})|^{\frac{1}{2}},$$

where $|\cdot|$ denotes the determinant and $I(\boldsymbol{\theta})$ is the expected Fisher information matrix with i, j-th element:

$$I_{ij}(\boldsymbol{\theta}) = \mathrm{E}\left[\frac{\partial^2}{\partial\theta_i\partial\theta_j}\log p(\mathbf{y}|\boldsymbol{\theta})\right].$$

Jeffreys' prior provides an automated procedure for finding an uninformative prior for any parametric model. Moreover, the Jeffreys prior is invariant under re-parameterisation of the parameter θ for in univariate and multivariate cases. However, one disadvantage of the Jeffreys prior is that the prior is improper for many models and it can cause improper posterior distributions. The prior $p(\theta)$ is said to be improper if:

$$\int p(\theta)d\theta = \infty.$$

For example, a uniform prior distribution on the real line is an improper prior.

Another well-known prior is the reference prior [17] that does not express personal beliefs. Reference priors are estimated by maximising some measure of distance or divergence between the posterior distribution and the prior distribution. There are a few divergence measures can be possibly chosen from, for example the Kullback-Leibler divergence [18] or the Hellinger distance. By maximising the divergence,

the data can result in a maximum impact on the posterior estimates. The reference priors and Jeffreys priors are equivalent for one dimensional parameters. However, for multidimensional parameters, they are different.

Conjugate Prior A prior is called a conjugate prior if the posterior distribution is from the same family of densities as the prior distribution. The prior and posterior are called conjugate distributions. For example, the Gaussian family is conjugate to itself with respect to a Gaussian likelihood function. Other example is that a beta prior is a conjugate prior for the binomial distribution. All members of the exponential family of distributions have conjugate priors. The conjugate prior simplifies the computation of the posterior distribution.

2.1.2.2 Posterior Distributions

The posterior distribution reflects our belief of the parameter after seeing the data by taking account both the prior information about θ and the data through the likelihood function.

There are several statistics used to characterise the properties of the posterior distribution:

- The posterior mode

  A mode refers to a local maximum of the probability density function. The distribution is unimodal if it has only one mode and multimodal if it has more than one mode. The posterior mode is a value that maximises the posterior distribution:

  $$\hat{\theta} = \arg\max_{\theta} p(\theta|\mathbf{y}).$$

  For a continuous posterior distribution, the posterior mode is calculated by computing derivative of the posterior distribution and setting it to zero.

- The posterior mean

The posterior mean is the expected value of the posterior distribution:

$$\hat{\theta} = \mathrm{E}(\theta|\mathbf{y}) = \int \theta\, p(\theta|\mathbf{y}) d\theta.$$

When the distribution has a heavy tail, the posterior mean can be far away from most of the probability.

- The posterior median

  The posterior median is the value such that half of the posterior distribution is below it and other half of the distribution is above it:

$$\hat{\theta} : p(\theta \geq \hat{\theta}|\mathbf{y}) = p(\theta \leq \hat{\theta}|\mathbf{y}) = \frac{1}{2}.$$

  The posterior median frequently has to be computed numerically. Posterior median is invariant to transformations of single coordinates.

2.1.2.3 Example

Suppose we survey n people on whether they like the color red or prefer other colors. Let θ be the probability one person prefers color red, we assume that all people are independent and θ is constant for all people. Let Z_i be the preference for person i, then it is a Bernoulli distribution:

$$Z_i|\theta \sim \text{Bernoulli}(\theta).$$

Now define $Y = \sum_i^n Z_i$ to be the total number of people prefer color red, it becomes a Binomial distribution:

$$Y|\theta \sim \text{Binomial}(n, \theta),$$

with Binomial likelihood expressed as:

$$p(y|\theta) = \binom{n}{y} \theta^y (1-\theta)^{n-y},$$

with mean $\mathrm{E}(Y) = n\theta$. Figure 2.1 shows four sample binomial distributions with $n = 100$ and $\theta = [0.1, 0.3, 0.5, 0.9]$. Next, we construct the Jeffreys uninformative

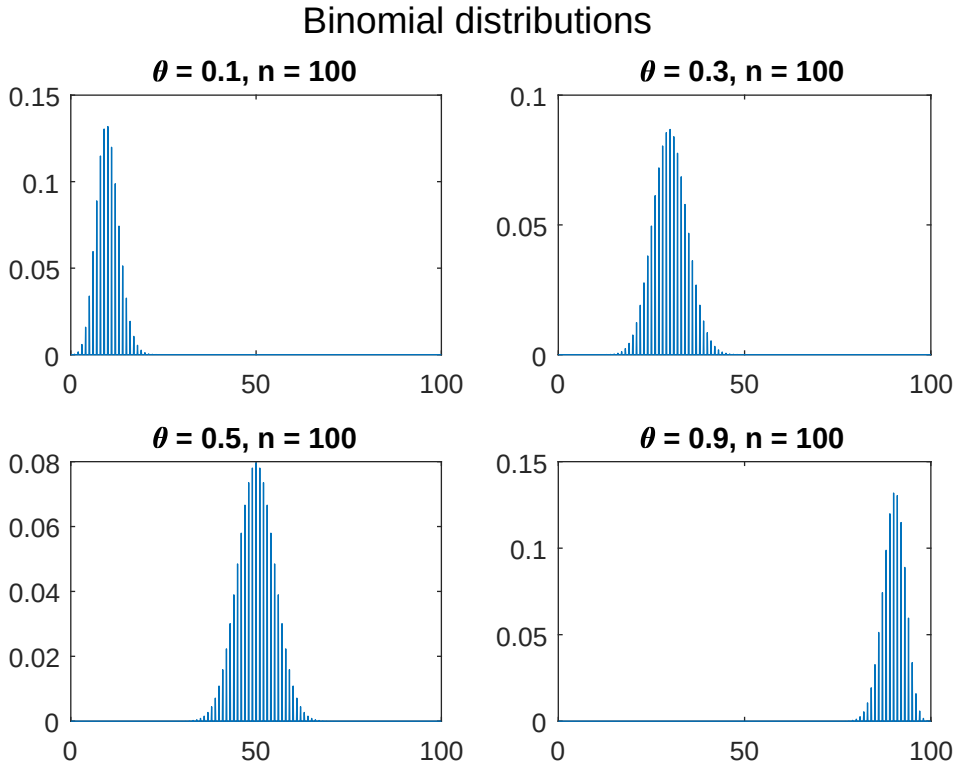

FIGURE 2.1: Plots of Binomial distributions with $n = 100$ and $\theta = [0.1, 0.3, 0.5, 0.9]$.

prior based on the binomial likelihood. Recall the Jeffreys prior (2.6) is:

$$p(\theta) \propto I(\theta)^{\frac{1}{2}},$$

where

$$\begin{aligned}
I(\theta) &= -\mathrm{E}_{y|\theta}\left[\frac{\partial^2}{\partial\theta^2}\log p(y|\theta)\right] \\
&= -\mathrm{E}_{y|\theta}\left[\frac{\partial^2}{\partial\theta^2}\left(\log\binom{n}{y} + y\log(\theta) + (n-y)\log(1-\theta)\right)\right] \\
&= -\mathrm{E}_{y|\theta}\left[\frac{\partial}{\partial\theta}\left(\frac{y}{\theta} - \frac{n-y}{1-\theta}\right)\right] \\
&= -\mathrm{E}_{y|\theta}\left[-\frac{y}{\theta^2} - \frac{n-y}{(1-\theta)^2}\right] \\
&= \frac{n}{\theta} + \frac{n-n\theta}{1-\theta} \\
&= \frac{n}{\theta(1-\theta)}.
\end{aligned}$$

Hence, the Jeffreys prior for the binomial likelihood is:

$$p(\theta) \propto I(\theta)^{\frac{1}{2}} \propto \theta^{-\frac{1}{2}}(1-\theta)^{-\frac{1}{2}},$$

which is in the form of beta distribution Beta$(\frac{1}{2}, \frac{1}{2})$.

More generally, if parameter $\theta \sim \text{Beta}(\alpha, \beta)$, the density function is described as:

$$p(\theta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}, \tag{2.7}$$

with mean $\frac{\alpha}{\alpha+\beta}$. Figure 2.2 shows beta probability density functions with different parameter values. When α and β have equal values, the distribution is symmetric. From (2.7) we see that when $\theta = \frac{1}{2}$, the data has the least impact on the posterior; when θ is near 0 or 1, the data has the most impact on the posterior. When $\alpha = \beta = \frac{1}{2}$, the Jeffreys prior puts more mass near 0 or 1 where the data has the strongest effect.

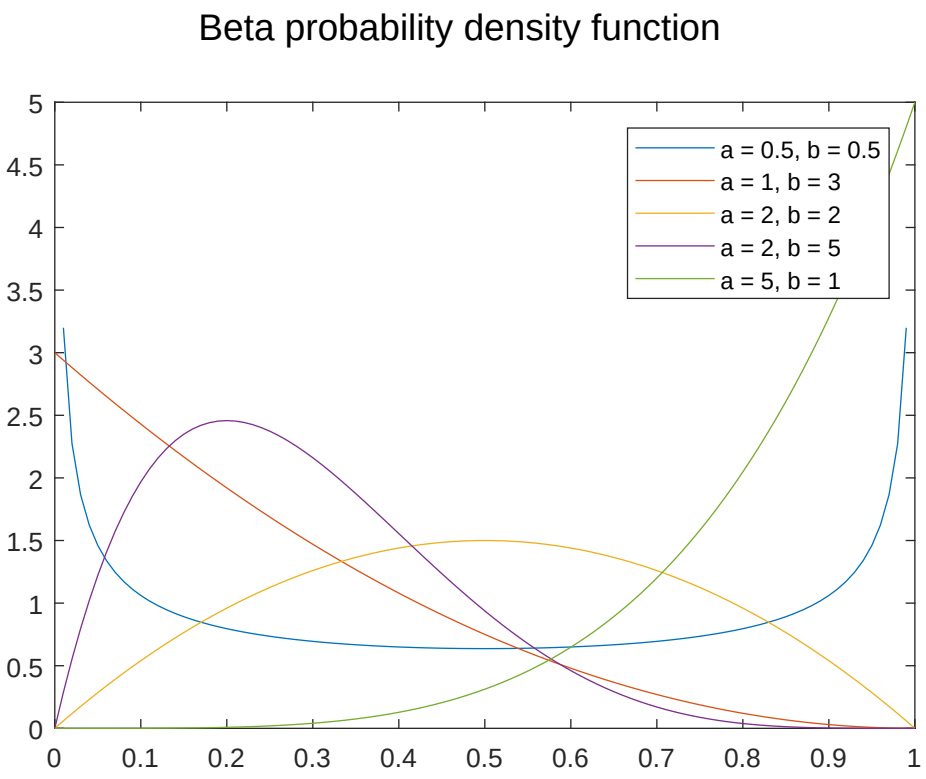

FIGURE 2.2: Plots of Beta probability density functions for five set of parameters: $\alpha = 0.5, \beta = 0.5$; $\alpha = 1, \beta = 3$; $\alpha = 2, \beta = 2$; $\alpha = 2, \beta = 5$; $\alpha = 5, \beta = 1$.

By using the Bayes' Theorem, the posterior distribution based on a Beta prior and Binomial likelihood is:

$$\begin{aligned} p(\theta|y) &= \frac{p(y|\theta)p(\theta)}{\int_0^1 p(y|\theta)p(\theta)d\theta} \\ &= \frac{\binom{n}{y}\theta^y(1-\theta)^{n-y}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}}{\int_0^1 \binom{n}{y}\theta^y(1-\theta)^{n-y}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}d\theta} \\ &= \frac{\theta^{y+\alpha-1}(1-\theta)^{n-y+\beta-1}}{\frac{\Gamma(\alpha+y)\Gamma(\beta+n-y)}{\Gamma(\alpha+\beta+n)}} \\ &\stackrel{d}{=} \text{Beta}(\alpha+y, \beta+n-y). \end{aligned}$$

Hence, the posterior distribution is a Beta distribution. From this example, we see that the posterior distribution belongs to the same family of prior distribution which is a beta distribution.

This example shows that if we start with a Beta(α, β) prior for parameter θ, after

data is observed, the posterior of θ is now a Beta$(\alpha + y, \beta + n - y)$ where extra data information has been incorporated into the prior belief, which results in an updated model.

Since the posterior distribution has a well-known beta distribution which also allows for calculation of some important quantities of interest. For example, the posterior mean is:

$$\bar{\theta} = \mathrm{E}(\theta|y) = \frac{\alpha + y}{\alpha + \beta + n}; \tag{2.8}$$

the posterior mode is derived by maximising the logarithm of posterior distribution:

$$\log p(\theta|y) = (\alpha + y - 1)\log(\theta) + (\beta + n - y - 1)\log(1 - \theta);$$

taking derivative of the logarithm function with respect to θ:

$$\frac{\partial}{\partial\theta}\log p(\theta|y) = \frac{\alpha + y - 1}{\theta} - \frac{\beta + n - y - 1}{1 - \theta} = 0;$$

and setting to zero to solve for the posterior mode:

$$\hat{\theta} = \frac{\alpha + y - 1}{\alpha + \beta + n - 2}. \tag{2.9}$$

For comparison, maximum likelihood estimator (MLE) is also computed for the model, which is solved by maximising likelihood function:

$$\hat{\theta}_{\mathrm{MLE}} = \arg\min_{\theta} \left\{ -\log L(\theta) \right\},$$

Noting that:

$$\begin{aligned} \log L(\theta) &= \log \binom{n}{y} + y\log(\theta) + (n - y)\log(1 - \theta) \\ \frac{\partial}{\partial\theta} - \log L(\theta) &= \frac{\partial}{\partial\theta}\left(-\log \binom{n}{y} - y\log(\theta) - (n - y)\log(1 - \theta) \right) \\ &= -\frac{y}{\theta} + \frac{n - y}{1 - \theta} = 0. \end{aligned}$$

The MLE of θ is:

$$\hat{\theta}_{\text{MLE}} = \frac{y}{n}, \tag{2.10}$$

which is the sample proportion. Instead of maximising the Bayesian analogue of the likelihood, the Bayesian estimation maximises the probability of the posterior distribution conditional upon the observed data, which allows for an explicit method to incorporate prior information into the data. The forms of Bayesian point estimates are similar to the MLE but with extra number of people and extra number of people select red color. The posterior mean in (2.8) is between MLE (2.10) and the prior mean $\alpha/(\alpha+\beta)$. When $\alpha = \beta = 1$, the prior is the uninformative prior with uniform distribution which implies there is no prior information of parameters available, then the posterior mode (2.9) is simply obtained from the observed data itself and it equals the MLE. Suppose α and β are fixed, in the extreme cases where none of the people choose red, i.e. $y = 0$, the posterior mean and mode estimates approach zero as n goes to infinity; and where all people choose red, i.e. $y = n$, the posterior estimates approach one as n goes infinity. The MLE in these situations produce exactly zero and one respectively.

As an illustration, suppose the Jeffreys prior Beta$(\frac{1}{2}, \frac{1}{2})$ is used for the beta-binomial model, the survey is conducted on $n = 10$ people and $y = 4$ of them choose color red as their preference. Figure 2.3 shows the plot of a beta prior distribution with a binomial likelihood.

2.1.3 Bayes Estimators

Let $\hat{\theta}$ be an estimator of θ and $L(\theta, \hat{\theta})$ be a loss function, then the Bayes risk is defined as:

$$\text{E}(L(\theta, \hat{\theta})) = \int L(\theta, \hat{\theta}) p(\mathbf{y}|\theta) d\mathbf{y}. \tag{2.11}$$

A Bayes estimator is an estimator that minimises the Bayes risk among all estimators. In other words, a Bayes estimator is the estimator that minimises the posterior expected loss $\text{E}(L(\theta, \hat{\theta})|\mathbf{y})$.

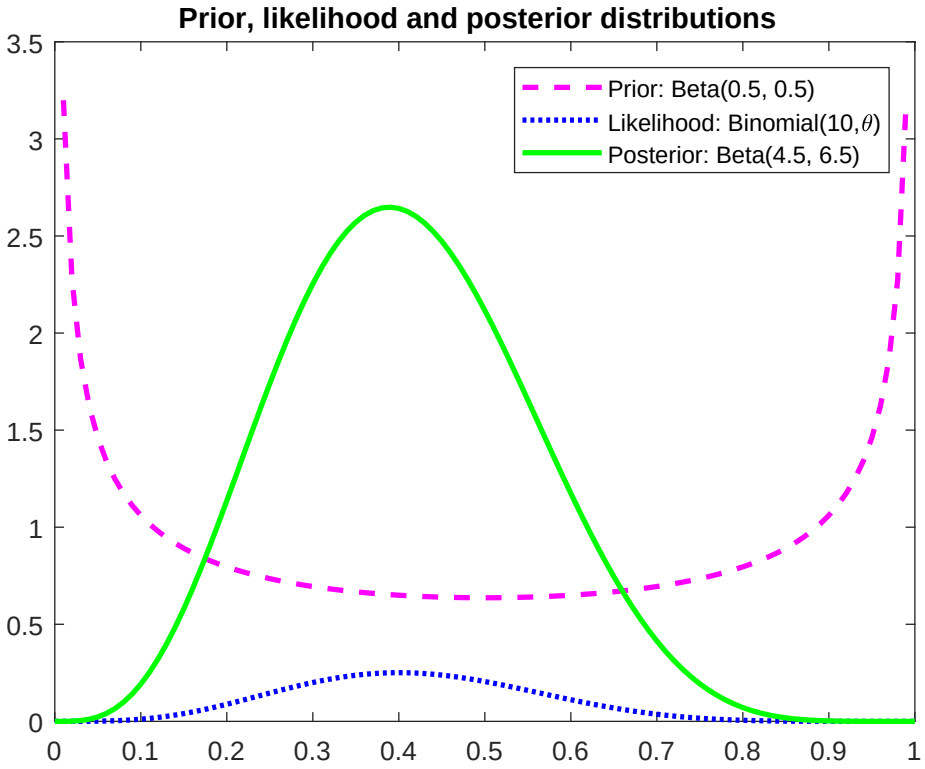

FIGURE 2.3: A Beta-Binomial model with beta prior distribution, binomial likelihood and beta posterior distribution

If the loss function is the squared error loss $L(\theta, \hat{\theta}) = (\theta - \hat{\theta})^2$, the risk function becomes mean squared error, i.e. $R(\theta, \hat{\theta}) = \mathrm{E}_\theta[(\hat{\theta} - \theta)^2]$, then the Bayes estimator of the unknown parameter is:

$$\hat{\theta}(y) = \int \theta \, p(\theta|y) d\theta = \mathrm{E}(\theta|y),$$

which is the mean of the posterior distribution.

2.1.4 Credible Intervals

After the estimate of the parameter is obtained, one may wish to find a high probability interval for the parameter. The Bayesian credible intervals or Bayesian confidence interval is the range of θ values with posterior probability $1 - \alpha$. Given

the posterior distribution $p(\theta)$, a credible interval C for parameter θ is defined as:

$$p(\theta \in C|\mathbf{y}) = \int_C p(\theta|\mathbf{y})d\theta.$$

For example, when $\alpha = 0.05$, the 95% credible interval is the interval C so that $\int_A p(\theta|\mathbf{y})d\theta = 0.95$.

Since posterior probability is a conditional probability conditioned on randomly observed data, so a credible interval of the posterior probability can be computed from the random variable to express its amount of uncertainty. The mode of the posterior distribution becomes the estimate of the parameter, and "probability intervals" which refers to the Bayesian analog of confidence intervals, can be calculated through the standard procedure.

The Bayesian credible interval does not have the property of including the true parameter with probability $1 - \alpha$ when sampling from the same parameter θ. Instead, it has probability $1 - \alpha$ containing the true parameter given the data.

2.1.5 Posterior Predictive Distribution

Given a fixed θ, data $\mathbf{Y}$ follows a distribution $p(\mathbf{y}|\theta)$. Since the true value of θ is unknown, all possible values of θ should be taken into account to obtain a better understanding of the distribution of $\mathbf{Y}$. The prior distribution $\pi(\theta)$ has accounted for the uncertainty in θ and therefore to predict the values of a future observation y^{new}, the posterior predictive distribution is described as:

$$p(y^{\text{new}}|\mathbf{y}) = \int_\theta p(y^{\text{new}}|\theta) \cdot \pi(\theta|\mathbf{y})d\theta \tag{2.12}$$

Instead of plugging in a single best estimate $\hat{\theta}$ into the data distribution $p(y^{\text{new}}|\hat{\theta})$, the Bayesian posterior predictive distribution provides a wider predictive distribution that accounts for the uncertainty in estimating θ.

2.1.6 Bayes Factors

The Bayesian approach to hypothesis testing is an alternative to the traditional frequentist approach and Bayesian model comparison is a way of model selection based on Bayes factors. It uses a Bayes factor to compare multiple models to determine which model fits data better [16].

For model selection problem, we need to choose between two models, on the basis of observed data $\mathbf{y}$. The plausibility of the two different models, say M_0 and M_1, parameterised by model parameters θ_0 and θ_1 is assessed by the Bayes factor:

$$BF = \frac{p(\mathbf{y}|M_0)}{p(\mathbf{y}|M_1)} = \frac{\int p(\theta_0|M_0)p(\mathbf{y}|\theta_0, M_0)d\theta_0}{\int p(\theta_1|M_1)p(\mathbf{y}|\theta_1, M_1)d\theta_1},$$

where $p(\mathbf{y}|M_0)$ is the marginal likelihood of data in Model 0. Therefore, the Bayes factor is a weighted average likelihood ratio, where the weights are based on the prior distribution specified for the hypotheses. It can also be viewed as the posterior odds in favor of the hypothesis divided by the prior odds in favor of the hypothesis. For example, if the Bayes factor is 3, then it means that the data favor model 0 over model 1 with $3:1$ odds.

Bayes factors can be defined only if the marginal distribution of data $\mathbf{y}$ is proper. If the marginal is improper, then it cannot be computed. However, there exist other ways to approximate to the integral such as the Laplace-Metropolis estimator [19, 20].

2.1.7 Markov Chain Monte Carlo

The main task of the Bayesian analysis is to compute the posterior density for the parameters. The difficulties lie in the denominator of the posterior distribution where the complex multidimensional integrals may need to be computed. Many techniques are available to perform the numerical integration. For example, the Simpson's rule [21] and the mid-point rule [22]. Modern techniques such as the Laplace approximation [23, 24] and Monte Carlo integration [25] have also gained

a lot of interest. The main focus of Bayesian computing today has been on the Markov chain Monte Carlo (MCMC) approach for sampling from the probability distribution and approximate the posterior distribution of a parameter of interest.

The MCMC sequentially samples parameter values to form a Markov chain where the stationary distribution is the desired joint posterior distribution. The popular algorithms for Bayesian MCMC are the Metropolis-Hastings algorithm [26, 27] and the Gibbs sampler [28].

2.1.7.1 Metropolis-Hastings Algorithm

Metropolis-Hastings (MH) algorithm is one of the Markov chain Monte Carlo methods for sampling from complex, non-standard probability distributions. It is particularly useful in sampling from multivariate distributions of high dimension. Under Bayesian framework, when it is difficult to evaluate the denominator analytically, MH can be used to draw samples from an intractable posterior distribution and then to calculate integrals or any numerical results of interest.

Suppose $p(\theta|\mathbf{y})$ is the target distribution we want to sample from and it is known up to a constant multiplier. The transition kernel is $Q(\theta'|\theta)$ which is a conditional distribution that represents the probability of moving from a current position θ to a new position θ'. If the distribution Q is symmetric, then $Q(\theta'|\theta) = Q(\theta|\theta')$ and it is also known as Metroplis algorithm. The MH method works as follows:

- Initialise $\theta^{(0)}$
- For steps $k = 1, 2, \ldots$:
  - Generate θ^* from $Q(\theta^*|\theta^{(k)})$.
  - Generate u from a uniform distribution $U(0, 1)$
  - Compute the MH acceptance ratio:

$$\alpha = \min\left(1, \frac{p(\theta^*|\mathbf{y})Q(\theta^{(k)}|\theta^*)}{p(\theta^{(k)}|\mathbf{y})Q(\theta^*|\theta^{(k)})}\right) = \min\left(1, \frac{p(\mathbf{y}|\theta^*)p(\theta^*)Q(\theta^{(k)}|\theta^*)}{p(\mathbf{y}|\theta^{(k)})p(\theta^{(k)})Q(\theta^*|\theta^{(k)})}\right)$$

– If $u \leq \alpha$, set $\theta^{(k+1)} = \theta^*$; otherwise set $\theta^{(k+1)} = \theta^{(k)}$.

Repeat the above procedure for sufficient iterations until a sequence of points are generated to approximate the posterior distribution.

MH does not involve any derivation analytically because it only requires the joint distribution. In addition, it works well in high-dimensional spaces.

2.1.7.2 Gibbs Sampler

The idea in Gibbs sampling [28] is to generate posterior samples by iterating through each variable or a block of variables to sample from its conditional distribution while the remaining variables are remain same as their current values. The Gibbs sampling works as follows:

- Specify initial starting values $\{\theta_1^{(0)}, \theta_2^{(0)}, \ldots, \theta_p^{(0)}\}$
- Draw $\theta_1^{(k)}$ from $p(\theta_1|\theta_2^{(k-1)}, \ldots, \theta_p^{(k-1)}, \mathbf{y})$
- Draw $\theta_2^{(k)}$ from $p(\theta_2|\theta_1^{(k)}, \theta_3^{(k-1)}, \ldots, \theta_p^{(k-1)}, \mathbf{y})$

  $\vdots$
- Draw $\theta_p^{(k)}$ from $p(\theta_p|\theta_1^{(k)}, \theta_2^{(k)}, \ldots, \theta_{p-1}^{(k-1)}, \mathbf{y})$
- Update $\boldsymbol{\theta}$ until convergence.

The posterior distribution is not sampled directly, samples are simulated based on all the posterior conditionals, one random variable at a time. Since the initial value is for the Gibbs sampling is selected randomly, all samples simulated at early iterations may not be representative of the actual posterior distribution. On the other hand, MCMC guarantees that the stationary distribution of the samples is our target joint posterior distribution. Therefore, MCMC algorithms are generally run for a large number of iterations until that convergence to the target posterior is achieved. Because samples from the early iterations do not necessarily from our target posterior, we generally remove those samples. The discarded iterations are often referred to as the "burn-in" samples.

2.2 Linear Regression

Linear regression is a class of statistical models that provides a study of relationships between two variables. In this chapter, we will cover simple and multiple linear regression where both variables are continuous variables, as well as logistic regression which is used to characterise the discrete response variable.

2.2.1 Linear Regression Model

Linear regression is a widely used approach in statistics for modeling the relationship between the dependent variable and independent variables. The main usage of the linear regression model includes examining which of the independent variables are related to the dependent variables and predicting the value of the dependent variable given the independent variables.

2.2.1.1 Simple Linear Regression

In the simplest case of a regression model with only one independent variable, we call this simple linear regression. The simple linear regression has the following form:

$$Y = \beta_0 + \beta_1 X + \epsilon, \tag{2.13}$$

where Y is a dependent variable (also known as response variable, regressand), X is an independent variable (also known as explanatory variable, covariate, predictor variable, regressor), β_0 and β_1 are regression parameters, and ϵ is an error term (also known as noise variable, disturbance term). For the linear regression model, it is assumed that the dependent variable Y is a continuous variable and independent variable X can be either continuous or discrete.

Simple linear regression in equation (2.14) connects Y and X through regression parameters β_0 and β_1. It represents a straight line with β_0 corresponding to the intercept and β_1 indicating the slope of the straight line. If n data points are given

for the independent variable X with values $x_1, \ldots, x_n$, then Y also takes n values $y_1, \ldots, y_n$. Given data, the aim is to find the best straight line that describe the data.

2.2.1.2 Multiple Linear Regression

A more general form of the regression model is the multiple linear regression where it contains more than one predictor. Suppose there are p predictors in the model, then the multiple linear regression model can be expressed as:

$$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p + \epsilon. \tag{2.14}$$

In multiple linear regression, data are represented by a data matrix. Given the data $\{(y_i, \mathbf{x}_i),\ i = 1, \ldots, n\}$, the multiple linear regression can also be expressed in matrix form:

$$y_i = \mathbf{x}_i^\top \boldsymbol{\beta} + \epsilon_i,\ i = 1, \ldots, n, \tag{2.15}$$

where $\mathbf{x}_i = (1, x_{i1}, \ldots, x_{ip})$ and $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^\top$ is a vector of regression coefficients.

2.2.1.3 General assumptions of linear regression

The usual assumptions of a linear regression model include the following:

- The dependent variable Y is a continuous variable
- The relationship between the independent and dependent variable is linear, i.e. Y is a linear function of X.
- The error terms ϵ_i are independent and identically distributed as per a $N(0, \sigma^2)$ with some finite variance σ^2.
- The error terms are independent of the predictors.
- There is little or no multicollinearity in the data.

The above assumptions are generally used in analysing real data although they may not often be true for most datasets in real-world situation.

2.2.1.4 Least Squares Estimation

The most common and popular way of fitting a linear regression model is the least squares method. The ordinary least squares (OLS) estimator is achieved by minimising the residual sum of squares (RSS) between the observations:

$$\mathrm{RSS}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left(y_i - \mathbf{x}_i^\top \boldsymbol{\beta}\right)^2.$$

It can also be written in matrix form:

$$\mathrm{RSS}(\boldsymbol{\beta}) = \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2,$$

where $\|\cdot\|_2$ is the ℓ_2-norm, which is also known as Euclidean distance. When $\mathbf{X}$ is full rank, i.e. $\mathbf{X}^\top\mathbf{X}$ is nonsingular, the least squares estimators $\hat{\boldsymbol{\beta}}_{\mathrm{LS}}$ is given by:

$$\hat{\boldsymbol{\beta}}_{\mathrm{LS}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}. \tag{2.16}$$

The least squares estimator is an unbiased and consistent estimator. It in fact corresponds to the maximum likelihood estimation when the noise variable follows normal distributions with equal variance. If $\mathbf{X}$ is not of full rank and $\mathbf{X}^\top\mathbf{X}$ is not invertible, then there is no unique solution for $\hat{\boldsymbol{\beta}}_{\mathrm{LS}}$.

2.2.1.5 Parameter Estimation and Prediction

To quantify how "good" a particular estimate is, there are two main aspects:

i) whether the estimate is close to the "true coefficients";

ii) whether the estimate calculated from the available data fit well for the future data.

Mean Squared Error To address the first aspect, one of the common approaches, the mean squared error (MSE), is taken into consideration since it is the squared distance between an estimator $\hat{\boldsymbol{\beta}}$ and the true underlying parameter $\boldsymbol{\beta}$:

$$\mathrm{MSE}(\hat{\boldsymbol{\beta}}) = \mathrm{E}_{\boldsymbol{\beta}}\left[(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^\top(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})\right].$$

Prediction Error An estimate that provides a good result of fitting the current data does not guarantee its good performance in fitting the future data. Therefore, to address the second quantity, the mean squared prediction error (PE) can be introduced to measure how well the estimate fits new data. The prediction error for a given point is:

$$\begin{aligned}\mathrm{PE}(\mathbf{x}) &= \mathrm{E}_{Y|\mathbf{X}=\mathbf{x}}\left[\left(Y - \hat{f}(\mathbf{X})\right)^2 |\mathbf{X} = \mathbf{x}\right] \\ &= \mathrm{Bias}^2\left(\hat{f}(\mathbf{x})\right) + \mathrm{Var}\left(\hat{f}(\mathbf{x})\right) + \sigma_\epsilon^2, \end{aligned} \tag{2.17}$$

where $\hat{f}(\mathbf{x}) = \mathbf{x}^\top\hat{\boldsymbol{\beta}}$ is the predictive function of a linear predictor $f(\mathbf{x}) = \mathbf{x}^\top\boldsymbol{\beta}$, and σ_ϵ^2 is an irreducible error which is a measure of how much noise is in the data. The error cannot be reduced no matter how good the model is created because there is always certain noise in the data.

As is shown by the formula, the prediction error is the expected value of the squared distance between the predictor for a certain value and the true value. It suggests that if $\hat{f}$ is a good model, it should fit well to the new data too.

Bias-Variance Trade-off The prediction error is a measurement of the quality of an estimator that consists of a bias term and a variance term. If more parameters are included in the model, then the model becomes more complex and therefore reduces the bias at the expense of an increase in variance. Conversely, if bias is sacrificed a little with less parameters, the variance can be substantially reduced which may lead to an overall reduction in the prediction error. This is known as the bias-variance trade-off.

In terms of the linear regression model, the prediction error can be expressed as:

$$\begin{aligned} \text{PE} =& \text{E}\left((\mathbf{y}-\hat{\mathbf{y}})^\top(\mathbf{y}-\hat{\mathbf{y}})\right) \\ =& \left(\text{Bias}(\mathbf{X}\boldsymbol{\beta}, \mathbf{X}\hat{\boldsymbol{\beta}})\right)^2 + \text{Var}(\epsilon) \\ =& (\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})^\top \text{cov}(\mathbf{X})(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}) + n\sigma_\epsilon^2. \end{aligned} \tag{2.18}$$

Although the least squares estimator is an unbiased estimate with minimum variance, the estimate is unconstrained and the variance can be large. The least squares method never return an estimate with zero coefficients. It is therefore possible to sacrifice some bias in order to reduce the variance. Moreover, with the all predictors included in the model, it can be difficult to interpret the regression model.

2.2.2 Logistic Regression

Generalised linear models [29] are a larger class of models compared to conventional general linear model which normally refers to the linear regression models.

For a linear regression model, we have assumed that the dependent variable Y is a continuous variable and independent variables X can be either continuous or discrete. However, response variable Y is not limited to continuous variable only in many practical situations. For example, response variables can be continuous variables defined on certain interval, binary variables, discrete counts, rates, proportions, etc. Generalised linear models are able to model such response variables.

A logistic regression is one of the generalised linear models where Y is a binary or dichotomous response variable coded as 1 for success and 0 for failure rather than a continuous variable. When the outcome is binary, logistic regression is a statistical method estimating the probability of success when there are one or more independent variables.

For logistic model, the distribution of Y_i is assumed to follow a Bernoulli distribution $Y_i \sim \text{Bernoulli}(\pi_i)$, where π_i is the probability of success for data i. The

logistic regression is regressing for the probability of a categorical outcome and the model has the following form:

$$\pi_i = P(Y_i = 1|\mathbf{X}_i = \mathbf{x}_i) = \frac{e^{\beta_0 + \mathbf{x}_i^\top \boldsymbol{\beta}}}{1 + e^{\beta_0 + \mathbf{x}_i^\top \boldsymbol{\beta}}}.$$

Notice the probability of failure is $P(Y_i = 0|\mathbf{X}_i = \mathbf{x}_i) = 1 - \pi_i$. It can also be expressed as the log odds of probability of success as shown below:

$$\text{logit}(\pi_i) = \log\left(\frac{\pi_i}{1 - \pi_i}\right) = \beta_0 + \mathbf{x}_i^\top \boldsymbol{\beta}.$$

Here, the link function of logistic regression is $\eta = \text{logit}(\pi) = \frac{\pi}{1-\pi}$ which is a logit link and odds of the dependent variable X is defined as:

$$\text{odds} = \frac{\pi}{1 - \pi} = e^{\beta_0 + \mathbf{X}^\top \boldsymbol{\beta}},$$

which equals the exponential function of the simple linear regression expression.

Recall that there exist a direct connection between Y_i and parameter $\boldsymbol{\beta}$ in the linear regression in (2.14), so the least squares estimate can be computed easily. The logistic regression does not has such desired property. To find estimates for the parameter $\boldsymbol{\beta}$, define the likelihood function to be:

$$L(\boldsymbol{\beta}|\mathbf{y}) = \prod_{i=1}^{n} \pi(x_i)^{y_i}(1 - \pi(x_i))^{1-y_i},$$

where

$$\pi(x_i) = \frac{e^{\beta_0 + \mathbf{x}_i^\top \boldsymbol{\beta}}}{1 + e^{\beta_0 + \mathbf{x}_i^\top \boldsymbol{\beta}}}.$$

The maximum likelihood estimates are then computed by maximising log likelihood function. By setting the derivative of negative log likelihood to zero, we

obtain two following equations:

$$\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \frac{1}{1 + e^{-\beta_0 - \mathbf{x}_i^\top \boldsymbol{\beta}}}$$
$$\sum_{i=1}^{n} y_i x_i = \sum_{i=1}^{n} \frac{x_i}{1 + e^{-\beta_0 - \mathbf{x}_i^\top \boldsymbol{\beta}}}.$$

Since the above two equations are nonlinear of β_0 and $\boldsymbol{\beta}$, the maximum likelihood estimates need to be computed numerically. If there exist a solution to the likelihood equations, the solution is unique. However, in some extreme case, there may not exist a solution as the maximum value of the likelihood happens in some limit as regression coefficients $\boldsymbol{\beta}$ approaches infinity.

2.3 Non-Bayesian Penalised Regression

Model selection, parameter estimation and prediction have been three main focuses of the regression analysis and have gained great interest. Besides the least squares estimation introduced previously, there are many other parameter estimation techniques such as maximum likelihood estimation, penalised regression, Bayesian linear regression, etc. Some of techniques automatically provide solution to the variable selection problem by producing sparse estimates for the parameters. We will cover penalised regression techniques in this section and Bayesian linear regression in Section 2.4.

2.3.1 Subset Selection

Recall that the prediction error in (2.18) consists of two components: a squared bias term and a variance term. Although the least squares estimator produces an unbiased estimate with minimum variance, the variance can be large. By sacrificing unbiasedness, a better mean squared error can be achieved. For models with a large number of predictors, another potential issue is that the least squares

estimator fails to identify the most important variables among the full set. Therefore, it may be helpful to find a biased estimate with fewer predictors, but lower variance and better prediction accuracy.

When the data are collinear or irrelevant to the response variable, some of the predictor variables can be redundant and can be eliminated. Subset selection [2, 3] is a method of selecting a subset of columns of predictor matrix $\mathbf{X}$, identifying the best subset among many others and using the remaining ones to predict Y.

Since the variance of regression is increased with each additional coefficient added, fewer covariates will result in a lower variance. In addition, it accelerates the prediction process, simplifies the regression equation and provides a clear understanding of which covariates $\mathbf{x}$ are related in predicting y. Therefore, subset-selection regression is often used as a statistics procedure because of its variance reduction and simplicity.

However, there are several limitations of subset selection. Subset selection is a discrete process, indeed it is a binary process with the variables being either selected or eliminated. Selecting the wrong variables will lead to a wrong model and potentially lead to a large bias. It is also unstable; that is, if one data point $(y_i, \mathbf{x}_i)$ is removed, then the new selected variables can be completely different from the previous sets. Moreover, selecting only a few of the coefficients may lead to a biased estimate and may result in a large bias of the regression model. Another potential problem of subset selection is the fact that it is NP-hard (nondeterministic polynomial-time hard) [30] which means that it is infeasible to search for all possible subsets when the number of features is moderate to large.

Other common subset selection methods in statistics include stepwise regression, forward selection or elimination. For more details and further examples, see Section 2.3.1.1, 2.3.1.2, and 2.3.1.3.

2.3.1.1 Forward Selection

The forward selection method [4] starts with no variables in the model, and at each step it adds to the model the variable with the largest absolute correlation with the response Y or residual vector, continuing until none of the remaining variables are significant. This is a fast and simple method, but it can also be too greedy which means that some important but correlated predictors are likely to be not included in the model.

2.3.1.2 Backward Elimination

Similar to the forward selection, backward elimination starts with a full model where all predictors are included [5]. At each step, remove one predictor with the least significance given a chosen critical level. Keep removing predictors until all remaining variables are statistically significant. However, sometimes predictors are deleted from the model but can be significant when added to the final reduced models.

2.3.1.3 Forward Stagewise Selection

The forward stagewise regression [6] seeks to solve the greediness of forward selection by adding variables partially. Whereas forward selection finds the variable with the largest absolute correlation with the residual and adds this variable to the model, forward stagewise finds the variable with the largest absolute correlation with the residual, computes the coefficient of the residual for that variable, and updates its weight by adding the current coefficient to that variable. It continues updating until there is no predictor corrected with the residual. The drawback of forward stagewise regression is that it makes thousands of tiny steps until convergence and can be very inefficient.

2.3.1.4 Mixed Integer Optimisation

Most recently, a modern optimisation method called mixed integer optimisation (MIO) [31] has been proposed for the best subset selection problem in linear regression. It solves for high quality feasible solutions through MIO and is a discrete extension of first order continuous optimisation methods. By using warm starts to a MIO solver, it finds provably optimal solutions. It has demonstrated its tractability by solving problems with n in 1000s and p in 100s in minutes. Moreover, it has shown good performance in selecting sparse model with excellent predictive power through an empirical study.

Hastie et al. [32] further extended the simulations to report a more comprehensive comparison with other state-of-the-art methods. The MIO algorithm on average perform better than the Lasso under the settings of high signal-to-noise ratio, it has performed similarly compared with forward stepwise selection.

2.3.2 Ridge Regression

As discussed in Section 2.2.1.5, the ordinary least squares estimator provides the smallest variance among all linear unbiased estimates. Is it possible to sacrifice a little unbiasedness and find an estimate which not only produces smaller variance but also has smaller mean squared error than OLS? In this section, the idea of shrinkage is introduced.

In 1956, Stein [33] considered the problem of estimating the mean vector of only one observation from a multivariate normal distribution with identity covariance under quadratic loss function. It is reasonable to use that one single observation vector $\mathbf{y}$ as the estimate and it is proven that this estimate is inadmissible for dimension $p \geq 3$. Several years later, James and Stein [34] found a shrinkage estimator of the form:

$$\left(1 - \frac{p-2}{||\mathbf{y}||^2}\right)\mathbf{y}.$$

This estimate dominates OLS estimator for dimension $p \geq 3$. Later on, Sclove [35] adapted the idea of James-Stein estimator to the regression model. Driven by the idea of the James-Stein estimator, a competitor to the subset selection, the ridge regression [8], has then been proposed. The ridge regression penalises the size of regression coefficients in terms of minimising the fitted sum of squares. By introducing a little bias to the regression estimates, ridge regression substantially reduces in variance and decreases in prediction error.

Without loss of generality, we assume that $\mathbf{y}$ is centered to have mean zero so that there is no intercept term, i.e. $\hat{\beta}_0 = 0$, and x_{ij} are standardised with mean zero and unit length:

$$\sum_{i=1}^{n} \frac{x_{ij}}{n} = 0, \quad \sum_{i=1}^{n} \frac{x_{ij}^2}{n} = 1.$$

The ridge regression estimate $\hat{\boldsymbol{\beta}}_{\text{ridge}}$ is defined to minimise the penalised sum of squares:

$$\sum_{i=1}^{n} \left(y_i - \mathbf{x}_i^\top \boldsymbol{\beta}\right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2, \tag{2.19}$$

or expressed in the matrix form:

$$\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_2^2, \tag{2.20}$$

where $\lambda \geq 0$ is called the tuning parameter (also known as shrinkage parameter or regularisation parameter), which controls the amount of shrinkage. The second term in (2.19) is called the penalty term. The estimator of ridge regression is:

$$\hat{\boldsymbol{\beta}}_{\text{ridge}} = \left(\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}\right)^{-1} \mathbf{X}^\top \mathbf{y}.$$

Here matrix $\left(\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}\right)$ is always invertible provided λ is positive. When $\lambda = 0$, the ridge estimate is equivalent to ordinary least squares estimate as shown in (2.16). The ridge estimate approaches zero as λ goes to infinity. For λ values in between 0 and ∞, the ridge estimate does both fitting and shrinking. The standard approach to select a value of λ is to use the cross-validation method.

In the orthonormal case, where $\mathbf{X}^\top\mathbf{X} = \mathbf{I}$, the ridge estimate becomes:

$$\hat{\boldsymbol{\beta}}_{\text{ridge}} = (\mathbf{I} + \lambda\mathbf{I})^{-1}\mathbf{X}^\top\mathbf{y} = \frac{\hat{\boldsymbol{\beta}}_{\text{LS}}}{1+\lambda}.$$

This clearly indicates the shrinkage nature of the ridge regression since $\lambda \geq 0$.

To extend the ridge regression to a more general case, the generalised ridge regression is:

$$\begin{aligned}\hat{\boldsymbol{\beta}}_{\text{ridge}} &= \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\arg\min} \left\{ \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda\boldsymbol{\beta}^\top\boldsymbol{\Sigma}\boldsymbol{\beta} \right\} \\ &= (\mathbf{X}^\top\mathbf{X} + \lambda\boldsymbol{\Sigma})^{-1}\mathbf{X}^\top\mathbf{y}.\end{aligned}$$

It can be seen that the ridge regression is a continuous process because for any $\lambda \geq 0$, a different model is obtained. There are potentially infinite number of models according to various λ values. By change λ, a model is changed smoothly to another. Although adding the penalty term may introduce some bias, it reduces the variance by shrinking the coefficients toward zero which improves prediction accuracy. Moreover, if one data is deleted from data sets, the ridge estimate should be close to the previous one and therefore it is stable procedure.

The drawbacks of ridge regression is that it cannot pick up the relevant variables from a larger set of variables and it provides no clearer interpretation than the ordinary least squares because it does not set any coefficient to exactly zero.

2.3.3 Non-Negative Garrote

Subset selection is unstable and inaccurate. Ridge regression shrinks the coefficient and reduces the variance, but it hard to interpret the relationship between y and multiple X. The non-negative garrote method [7] has been proposed to perform both variable selection and parameter estimation.

Let $\hat{\beta}_j^{\text{LS}}$ be the least squares estimates for predictor j, the non-negative garrote estimate is derived by:

$$\min_{c_j} \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} c_j \hat{\beta}_j^{\text{LS}} x_{ij} \right)^2 \quad \text{subject to } c_j \geq 0, \sum_{j=1}^{p} c_j \leq \lambda.$$

The non-negative garrote predictor coefficients are $\hat{\beta}_j^{\text{NNG}} = c_j \hat{\beta}_j^{\text{LS}}$. As λ decreases, more c_j become zero and hence some of the coefficients $\hat{\beta}_j^{\text{NNG}}$ are eliminated while the rest are shrunken toward zero.

In 1995, Breiman [7] has shown that non-negative garrote performs better than subset selection in terms of the prediction accuracy and produces more zero coefficients than subset selection does. Furthermore, its accuracy is also competitive with ridge regression and ridge only wins when most of the true coefficients are small and non-zero. The other advantage of non-negative garrote is that it is scale invariant. It is also relatively stable, with stability being intermediate between subset selection and ridge regression.

However, the estimate of non-negative garrote depends on the OLS estimate. When the OLS estimate does not perform well in the correlated case or over-fit situation, the non-negative garrote may be affected as well.

2.3.4 The Lasso

Since the non-negative garrote idea has been introduced, more penalised regressions have been explored. One of the most popular technique called the least absolute shrinkage and selection operator (Lasso) [9] has been developed. It does both shrinking and selecting, and so it retains good features of subset selection and ridge regression.

The Lasso estimate $\hat{\boldsymbol{\beta}}_{\text{Lasso}}$ is derived by minimising the penalised sum of squares:

$$\sum_{i=1}^{n} \left(y_i - \mathbf{x}_i^\top \boldsymbol{\beta} \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j|, \tag{2.21}$$

or expressed in matrix form:

$$\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1, \tag{2.22}$$

where $\|\cdot\|_1$ is the ℓ_1-norm and the tuning parameter $\lambda \geq 0$ also controls the amount of shrinkage as in the ridge regression case.

The Lasso sets some of the coefficients to exactly zeros while ridge regression only shrinks coefficients toward zero. As an illustration, let's consider the case when

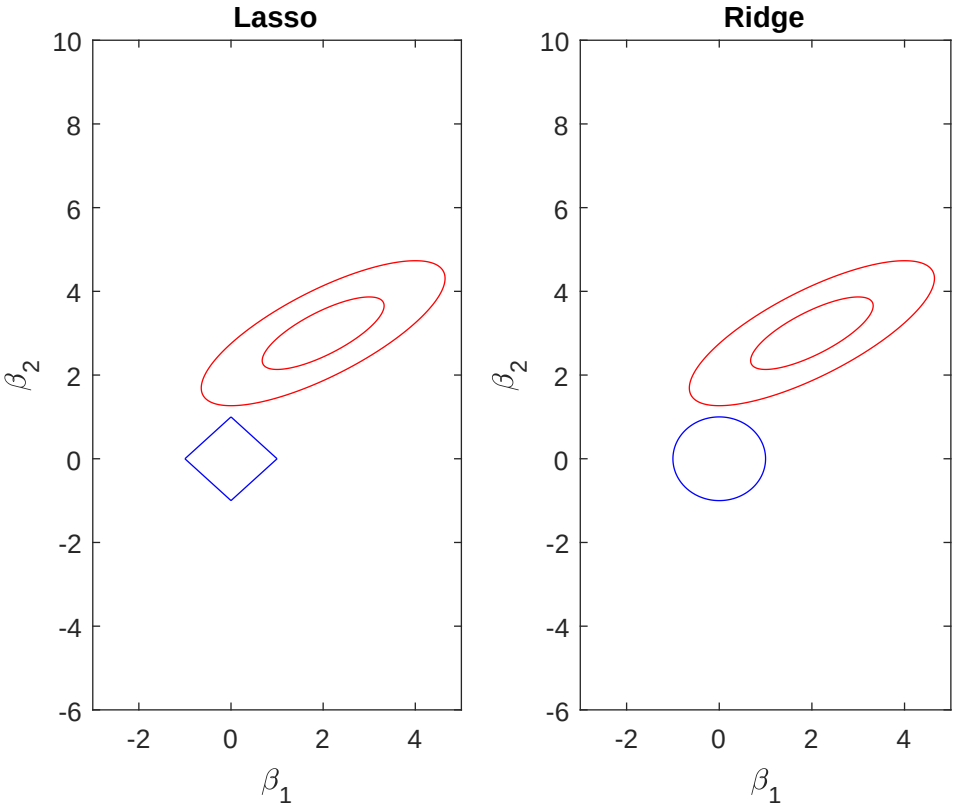

FIGURE 2.4: An illustration of the Lasso model versus the ridge model when $p = 2$.

there are two predictors, i.e. $p = 2$, the first term in (2.21) can be expressed as the quadratic form:

$$\left(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\right)^\top \mathbf{X}^\top\mathbf{X} \left(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\right)$$

which represents an elliptical contour. The constraint for the Lasso which is the second term in (2.21) is $|\beta_1| + |\beta_2|$ which is a rotated square while the constraint for the ridge regression in (2.19) is $\beta_1^2 + \beta_2^2$ which is a circle. The estimates can be

achieved in both cases only when contours touch the constraint region. As shown in Figure 2.4, for the Lasso, it can happen at the corner of the square which corresponds to the zero coefficients. But for the ridge, it can never happen at the corner. That is the reason why the Lasso can set some of the coefficients to zeros while the ridge cannot.

The Lasso is a continuous operation that shrink coefficients to exactly zero and thus selects subsets of predictors. With less predictors, it reduces the variance and provides a simpler model with a clear interpretation. It performs the best among non-negative garrote, ridge and subset selection when there are small to moderate number of moderate-sized effects. One advantage of the Lasso over non-negative garrote is that it does not rely an initial estimator, which could be the OLS estimator.

The Lasso remains the most popular and commonly used technique in the statistics literature. Many more penalised regressions then have been proposed such as the adaptive Lasso [11], the elastic net [36], the group Lasso [37], etc.

2.3.4.1 Least Angle Regression

One popular and most frequently used method for solving the Lasso and non-negative garrote estimate is the least angle regression (LARS) [38]. LARS is designed for small to moderate data and provides a selection of a subset of variables to include in the model and also their coefficient values.

The motivation of the LARS algorithm is from the forward stagewise regression introduced in Section 2.3.1.3. Instead of including variables at each step, the estimated parameters are increased in a direction equiangular to each parameter's correlation with the residual and makes optimally-sized leaps in the optimal directions. These directions are selected to make equal angles (i.e. equal correlations) with each variable currently in the model.

For simplicity, assume that explanatory variables are standardised to have zero mean and unit variance, and the response variable has zero mean. In detail, the LARS algorithm works as follows:

- Initiate with no variable in the model.
- Find the variable x_j that is most correlated with the residual. The variable that correlates with the residual the most is equivalent to the one that makes the least angle with the residual.
- Move in the same direction of this variable x_j until some other variable, say x_k, has as much correlation as x_j does. At this point, start to move in a direction such that the residual stays equally correlated with both x_j and x_k. It means that the residual makes equal angles with both variables. Keep moving until some other variable x_l becomes equally correlated with the residual.
- Repeat the above procedures until all predictors are included in the model.

2.3.4.2 Modified LARS For Lasso

LARS algorithm can also be used to solve Lasso. The modified LARS algorithm for the Lasso follows the same procedure as the LARS except for: if a coefficient hits zero, it will remain zero and thus drop the coefficient from the active set. Then the current joint least squares direction is recalculated and the same procedure is repeated.

The difference between the LARS and modified LARS for Lasso is that the LARS passes through zero while the Lasso reaches zero, and remains zero. With one modification, the LARS procedure can produce exactly the same set of Lasso solutions for all values of tuning parameters λ.

The LARS (Lasso) algorithm is efficient because it uses the same order of computation as the least squares fit based on p predictors, and LARS requires exactly

p steps to reach the full least squares estimates. The Lasso path may take more than p steps, although two procedures are usually quite similar. The LARS algorithm for Lasso is a very efficient way to obtain the solution to any Lasso problem, especially when n is smaller than p.

In conclusion, LARS uses least squares directions in the active set of variables; LARS for Lasso uses least square directions. If a variable hits zero, it is removed from the active set. However, the main disadvantage of LARS is that the algorithm performs relatively slow when p is large.

2.3.4.3 Coordinate-Wise Descent

When the data is of high dimension, the Lasso estimator can be solved using coordinate-wise descent method [39] which is simple, fast and stable. The coordinate-wise descent algorithms have been proposed to ℓ_1-penalised regression problem such as the Lasso for a class of convex optimisation problems. The key concept of the algorithms is that we can cycle through the variables and update the coordinate one at a time in multivariate functions [40].

The estimate of the coordinate-wise descent method is achieved by updating one coefficient while holding the other fixed and iteratively cycling until the coefficients converge. In order to applying coordinate descent, derivation of estimate at each step is needed. Without loss of generality, we assume that we minimise:

$$\frac{1}{2n}\sum_{i=1}^{n}\left(y_i-\sum_{j=1}^{p}x_{ij}\beta_j\right)^2+\lambda\sum_{j=1}^{p}|\beta_j|.$$

Then

$$\begin{aligned}
\hat{\beta}_j^{\text{Lasso}} &= \begin{cases} \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \sum_{k=1,k\neq j}^{p} x_{ik}\beta_k\right)x_{ij} - \lambda & \text{if } \hat{\beta}_j^{\text{Lasso}} > 0 \\ \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \sum_{k=1,k\neq j}^{p} x_{ik}\beta_k\right)x_{ij} + \lambda & \text{if } \hat{\beta}_j^{\text{Lasso}} < 0 \end{cases} \\
&= \begin{cases} \hat{\beta}_j - \lambda & \text{if } \hat{\beta}_j > 0 \text{ i.e. } \hat{\beta}_j > \lambda \\ \hat{\beta}_j + \lambda & \text{if } \hat{\beta}_j < 0 \text{ i.e. } \hat{\beta}_j < -\lambda \end{cases}, \quad \lambda > 0 \\
&= \text{sign}\left(\hat{\beta}_j\right)\left(|\hat{\beta}_j| - \lambda\right)_+,
\end{aligned}$$

where $\hat{\beta}_j$ is the least squares estimate. After several iterations, the estimate will converge and give us the estimate for the Lasso.

Although coordinate-wise descent is a simple technique, it is surprisingly efficient and scalable, and therefore it is competitive with LARS especially when there is a large number of predictors in the Lasso problem.

2.3.4.4 Alternating Direction Method of Multipliers

Another popular method of solving Lasso problem is the alternating direction method of multipliers (ADMM) [41, 42]. The idea of ADMM is to extend augmented Lagrangian and the method of multiplier [43] by considering the decomposability of dual ascent algorithm [44] to solve convex optimisation problems. Therefore, ADMM enjoys the benefits of both dual decomposition and augmented Lagrangian.

ADMM considers the optimisation problem by splitting variables into two parts, $\mathbf{x}$ and $\mathbf{z}$, and solves a problem of the form

$$\begin{aligned}
&\min \ f(\mathbf{x}) + g(\mathbf{z}) \\
&\text{subject to } \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{z} = \mathbf{c}.
\end{aligned} \tag{2.23}$$

The above problem can be rewritten using augmented Lagrangian

$$L_\rho(\mathbf{x}, \mathbf{z}, \mathbf{u}) = f(\mathbf{x}) + g(\mathbf{z}) + \mathbf{u}^\top(\mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{z} - \mathbf{c}) + \frac{\rho}{2}\|\mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{z} - \mathbf{c}\|_2^2, \tag{2.24}$$

where $\mathbf{u}$ is the dual parameter or Lagrange multiplier, and $\rho > 0$ is the penalty parameter. This optimisation problem can be solved using the ADMM by minimising $L_\rho(\mathbf{x}, \mathbf{z}, \mathbf{u})$ over $\mathbf{x}, \mathbf{z}, \mathbf{u}$ iteratively until convergence. More specifically, it works by first minimising the augmented Lagrangian over variable $\mathbf{x}$ with both $\mathbf{z}$ and Lagrangian multiplier $\mathbf{u}$ being fixed. Next, it minimises the augmented Lagrangian over $\mathbf{z}$ with updated $\mathbf{x}$ and $\mathbf{u}$ being fixed. Last step is to update $\mathbf{u}$ using gradient ascent algorithm and latest $\mathbf{x}$ and $\mathbf{z}$. Suppose $(\mathbf{x}^{(0)}, \mathbf{z}^{(0)}, \mathbf{u}^{(0)})$ is the initial value, for step $k = 1, 2, 3, \ldots$, the variables are then updated as

$$\begin{aligned}
\mathbf{x}^{(k+1)} &:= \arg\min_{\mathbf{x}} L_\rho(\mathbf{x}, \mathbf{z}^{(k)}, \mathbf{u}^{(k)}) \\
\mathbf{z}^{(k+1)} &:= \arg\min_{\mathbf{z}} L_\rho(\mathbf{x}^{(k+1)}, \mathbf{z}, \mathbf{u}^{(k)}) \\
\mathbf{u}^{(k+1)} &:= \mathbf{u}^{(k)} + \rho(\mathbf{A}\mathbf{x}^{(k+1)} + \mathbf{B}\mathbf{z}^{(k+1)} - \mathbf{c}).
\end{aligned} \tag{2.25}$$

ADMM can also be expressed in an easier way by replacing the dual variable $\mathbf{u}$ with $\mathbf{w} = \mathbf{u}/\rho$, and it is usually referred as ADMM in scaled form. The augmented Lagrangian in (2.24) can then be written as

$$L_\rho(\mathbf{x}, \mathbf{z}, \mathbf{w}) = f(\mathbf{x}) + g(\mathbf{z}) + \frac{\rho}{2}\|\mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{z} - \mathbf{c} + \mathbf{w}\|_2^2 - \frac{\rho}{2}\|\mathbf{w}\|_2^2,$$

and variables are updated as

$$\begin{aligned}
\mathbf{x}^{(k+1)} &:= \arg\min_{\mathbf{x}} f(\mathbf{x}) + \frac{\rho}{2}\|\mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{z}^{(k)} - \mathbf{c} + \mathbf{w}^{(k)}\|_2^2 \\
\mathbf{z}^{(k+1)} &:= \arg\min_{\mathbf{z}} g(\mathbf{z}) + \frac{\rho}{2}\|\mathbf{A}\mathbf{x}^{(k+1)} + \mathbf{B}\mathbf{z} - \mathbf{c} + \mathbf{w}^{(k)}\|_2^2 \\
\mathbf{w}^{(k+1)} &:= \mathbf{w}^{(k)} + \mathbf{A}\mathbf{x}^{(k+1)} + \mathbf{B}\mathbf{z}^{(k+1)} - \mathbf{c}.
\end{aligned} \tag{2.26}$$

One of many examples of ADMM is the Lasso regression. Recall in Section 2.3.4, the Lasso is introduced and it is of the form

$$\min \frac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1. \tag{2.27}$$

So ADMM for Lasso problem can be expressed as

$$\begin{aligned} &\min \ \frac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\alpha}\|_1 \\ &\text{subject to } \boldsymbol{\beta} - \boldsymbol{\alpha} = 0, \end{aligned} \tag{2.28}$$

and augmented Lagrangian is

$$L_\rho(\boldsymbol{\beta}, \boldsymbol{\alpha}, \mathbf{w}) = \frac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\alpha}\|_1 + \frac{\rho}{2}\|\boldsymbol{\beta} - \boldsymbol{\alpha} + \mathbf{w}\|_2^2 - \frac{\rho}{2}\|\mathbf{w}\|_2^2.$$

ADMM update for Lasso is

$$\begin{aligned} \boldsymbol{\beta}^{(k+1)} &:= (\mathbf{X}^\top\mathbf{X} + \rho\mathbf{I})^{-1}(\mathbf{X}^\top\mathbf{y} + \rho(\boldsymbol{\alpha}^{(k)} - \mathbf{w}^{(k)})) \\ \boldsymbol{\alpha}^{(k+1)} &:= S_{\frac{\lambda}{\rho}}(\boldsymbol{\beta}^{(k+1)} + \mathbf{w}^{(k)}) \\ \mathbf{w}^{(k+1)} &:= \mathbf{w}^{(k)} + \boldsymbol{\beta}^{(k+1)} - \boldsymbol{\alpha}^{(k+1)}, \end{aligned} \tag{2.29}$$

where $S(\cdot)$ is a soft-threholding operator:

$$\left[S_{\frac{\lambda}{\rho}}(\mathbf{x})\right]_j = \begin{cases} x_j - \frac{\lambda}{\rho} & \text{if } x_j > \frac{\lambda}{\rho} \\ 0 & \text{if } x_j = \frac{\lambda}{\rho} \\ x_j + \frac{\lambda}{\rho} & \text{if } x_j < \frac{\lambda}{\rho} \end{cases} \quad j = 1, 2, \ldots, p.$$

It is noted here for $\rho > 0$, $\mathbf{X}^\top\mathbf{X} + \rho\mathbf{I}$ is always invertible regardless of $\mathbf{X}$.

The advantages of ADMM include that minimisation problem is separated into sub-problems, and enjoys decomposability. In addition, it enjoys better convergence properties of augmented Lagrangian as it only requires mild assumptions to ensure convergence.

2.3.5 The Adaptive Lasso

Although Lasso shows good performance in terms of variable selection, the Lasso is not consistent in general. It is consistent only if it satisfy a nontrivial condition. The oracle property states that: let $\mathcal{A} = \{j : \beta_j^* \neq 0\}$ and $\hat{\boldsymbol{\beta}}(\delta)$ is the coefficient estimator produced by a fitting procedure δ, then δ is called an oracle procedure if it identifies the right subset model, and has the optimal estimation rate:

$$\sqrt{n}(\hat{\boldsymbol{\beta}}(\delta)_{\mathcal{A}} - \boldsymbol{\beta}_{\mathcal{A}}^*) \xrightarrow{d} N(\mathbf{0}, \boldsymbol{\Sigma}^*),$$

where $\boldsymbol{\Sigma}^*$ is the covariance matrix of the true subset model.

Zou [11] proposed a modified version of the Lasso, the adaptive Lasso by penalising the adaptive weights. The advantages of the adaptive Lasso include the oracle properties and near-minimax optimality. Let $\mathcal{A}_n^* = \{j : \hat{\beta}_j^{*(n)} \neq 0\}$, suppose that $\hat{\boldsymbol{\beta}}$ is a root-n-consistent estimator to $\boldsymbol{\beta}^*$ such as ordinary least estimator, pick a $\gamma > 0$ and define the weight vector $\hat{\mathbf{w}} = 1/|\hat{\boldsymbol{\beta}}|^\gamma$, then the adaptive Lasso estimates $\hat{\boldsymbol{\beta}}^{*(n)}$ are given by:

$$\hat{\boldsymbol{\beta}}^{*(n)} = \arg\min_{\beta} ||\mathbf{y} - \sum_{j=1}^{p} \mathbf{x}_j \beta_j||^2 + \lambda_n \sum_{j=1}^{p} \hat{w}_j |\beta_j|.$$

The optimisation is convex and therefore it does not have the multiple local minimal issue. In addition, the adaptive Lasso is an ℓ_1 penalisation method, so current algorithms for solving Lasso such as LARS can also be applied to the adaptive Lasso.

With a proper choice of λ_n, the adaptive Lasso satisfies both consistency in variable selection, i.e. $\lim_n P(\mathcal{A}_n^* = \mathcal{A}) = 1$, and asymptotic normality, i.e. $\sqrt{n}(\hat{\boldsymbol{\beta}}_{\mathcal{A}}^{*(n)} - \boldsymbol{\beta}_{\mathcal{A}}^*) \xrightarrow{d} N(\mathbf{0}, \sigma^2 C_{11}^{-1})$. The theory and methodology of the adaptive Lasso can also be extended to generalised linear models.

In conclusion, the adaptive Lasso proposed for simultaneous estimation and variable selection has oracle propetries by utilising the adaptively weighted ℓ_1 penalty and leads to a near-minimax-optimal estimator. It also enjoys the computational

advantage of the Lasso and performs well compared to other sparse modeling techniques such as the Lasso, the SCAD [45], and the nonnegative garotte.

2.3.6 Smoothly Clipped Absolute Deviation

A good penalty function should satisfies three properties: unbiasedness, sparsity and continuity. Fan and Li [45] proposed a continuous differentiable penalty function called smoothly clipped absolute deviation (SCAD) penalty:

$$p'_\lambda(\beta_j) = \lambda \left\{ \mathbb{1}(\beta_j \leq \lambda) + \frac{(a\lambda - \beta_j)_+}{(a-1)\lambda} \mathbb{1}(\beta_j > \lambda) \right\},$$

for some $a > 2$ and $\theta > 0$. The resulting estimator is:

$$\hat{\beta}_j^{\text{SCAD}} = \begin{cases} \text{sign}(\hat{\beta}_j)(|\hat{\beta}_j| - \lambda)_+, & \text{when } |\hat{\beta}_j| \leq 2\lambda \\ \left\{(a-1)\hat{\beta}_j - \text{sign}(\hat{\beta}_j)a\lambda\right\}/(a-2), & \text{when } 2\lambda < |\hat{\beta}_j| \leq a\lambda \\ \hat{\beta}_j, & \text{when } |\hat{\beta}_j| > a\lambda \end{cases}$$

Two unknown parameters λ and a can be computed using the cross-validation method.

A new unified algorithm for minimising penalised regression via local quadratic approximations (LQA) has also been proposed [45] . The penalty term is approximated by a quadratic function

$$p_\lambda(|\beta_j|) \approx p_\lambda(|\beta_{j0}|) + \frac{1}{2}\left\{p'_\lambda(|\beta_{j0}|)/|\beta_{j0}|\right\}(\beta_j^2 - \beta_{j0}^2),$$

for $\beta_j \approx \beta_{j0}$. This method reduces the computational burden, but one limitation of the approximation is that once a coefficient is shrunken to zero, it will stay at zero.

Hence, this method can be used to select variables as well as estimate coefficients and it provides confidence intervals for the estimated parameters. It produces sparse solutions, reduce bias and results in continuous solutions under certain conditions.

The new unified algorithm proposed for optimising penalised likelihood functions is widely applicable. It can be applied to various parametric models such as generalised linear models and robust regression models, as well as nonparametric model such as splines. The proposed estimators perform as well as the oracle procedure in terms of variable selection with proper choice of regularisation parameters. In addition, standard errors of the estimated parameters can be accurately estimated. Compare to subset selection, this approach is much faster and more effective. Compare to Lasso, SCAD penalty function performs better in selecting significant variables without creating excessive biases.

2.3.7 Local Linear Approximation Algorithm

Zou and Li [46] proposed a one-step sparse estimation procedure for nonconcave penalised likelihood models. The new unified approach which is based on local linear approximation (LLA) algorithm to the penalty is:

$$p_\lambda(|\beta_i|) \approx p_\lambda(|\beta_j^{(0)}|) + p'_\lambda(|\beta_j^{(0)}|)(|\beta_i| - |\beta_j^{(0)}|) \quad \text{for } \beta_j \approx \beta_j^{(0)}.$$

To calculate the un-penalised maximum likelihood estimate, we solve for the following equation:

$$\boldsymbol{\beta}^{(k+1)} = \arg\max \left\{ \sum_{i=1}^{n} l_i(\boldsymbol{\beta}) - n \sum_{j=1}^{p} p'_\lambda(|\beta_j^{(k)}|)\beta_j| \right\},$$

and stop the iterations when $\boldsymbol{\beta}^{(k)}$ converges.

One-step procedure for sparse estimates are also derived as:

$$\boldsymbol{\beta}^{(1)} = \arg\min \left\{ \frac{1}{2}||\mathbf{y} - \mathbf{X}\boldsymbol{\beta}||^2 + n \sum_{j=1}^{p} p'_\lambda(|\beta_j^{(0)}|)|\beta_j| \right\},$$

and it enjoys the oracle properties under some conditions.

The advantages of LLA include computational efficiency, and the approximation is numerical stable because it does not need to delete any small coefficient or choose

the size of perturbation. The difference between the LLA and LQA is that the $\boldsymbol{\beta}^{(k+1)}$ and the final estimates naturally produces a sparse representation.

2.3.8 Tuning Parameters

One of the core components in the penalised regression is the tuning parameter λ. It controls the amount of regularisation and can be chosen by various methods, such as cross-validation, generalised cross-validation and the Stein's unbiased risk estimate.

Selection of the tuning parameter is of great importance since it can have a large effect on the performance of the estimates. We will discuss several common approaches to estimate λ.

2.3.8.1 Cross validation

One of the popular and simple methods for estimating a tuning parameter λ is called K-fold cross-validation method. The procedure is:

- Divide the data into K partitions of roughly equal size.
- For each $k \in \{1, \ldots, K\}$, leave kth part out, fit the model to the rest $K-1$ partitions with parameter λ and compute the prediction error in predicting the kth part:
$$\sum_{i \in k^{\text{th}} \text{ part}} \left(y_i - \mathbf{x}_i^\top \hat{\boldsymbol{\beta}}_{-k}(\lambda) \right)^2,$$
where $\hat{\boldsymbol{\beta}}_{-k}(\lambda)$ is the estimate for the regression coefficients based on the $k-1$ data partitions only.
- Calculate the cross-validation error by taking the average prediction error:
$$\text{CV}(\lambda) = \frac{1}{K} \sum_{k=1}^{K} \text{E}_k(\lambda).$$

- Compute the cross-validation estimate of prediction error for different value of λ and pick the one with minimum CV(λ).

Normally K is set to 5 or 10. When $K = n$, it is also called leave-one-out cross validation.

2.3.8.2 Information criteria

Various information-type methods are also approaches to select the regularisation parameter λ. By varying λ, the information criteria take all models with different λ values into consideration and return the best fitted model according to some criteria such as Mallow's Cp, the Akaike information criterion (AIC) [47], the Bayesian information criterion (BIC) [48], the symmetric Kullback information criterion (KIC) [49] and the minimum message length (MML) principle [50].

AIC The AIC is defined as:

$$-\ln L + p,$$

where L is the likelihood function given the data and p is the number of estimated parameters. It is proportional to the goodness of fit and penalises the number of parameters estimated, so the model with the lowest AIC value is the preferred model.

The limitation of AIC is that the penalty term of the number of parameters does not involve the sample size which implies that it is more likely to over-fit the model. AIC is inconsistent because it does not select the true model with probability 1 as the sample size grows to infinity.

AICc The poor behavior of the asymptotic approximation on which the AIC is based when n is small has lead to the proposal of the bias corrected version of AIC which is known as the corrected AIC (AICc) [51]. The AICc is defined as

$$-\ln L + \frac{pn}{n-p-1},$$

where n is the sample size and p is the number of parameters. Since it is basically AIC with a heavier penalty on the extra parameters, it converges to AIC when the sample size n is large.

BIC The BIC is given by:

$$-\ln L + \frac{p}{2}\ln(n).$$

The model with the minimum value of BIC is chosen as the preferred one. It is proven that as long as the true model is in the set of candidate models, BIC will select it with probability approaching one as $n \to \infty$ and therefore it is a consistent estimator. However, it is more likely to select a model with fewer parameters and therefore under-fit the model.

MML Minimum message length is originally a statement of information theory which suggests that the best model of the observed data is the one with the largest overall compression. It then has been applied as a criterion to linear regression problem [52] and has been developed as a method of model comparison by giving each model a score. MML is shown to be invariant and statistically consistent. Two recent examples of linear regression criteria based on MML are MMLu and MMLg [53].

2.4 Bayesian Linear Regression

An important class of alternative penalised regression methods are the Bayesian penalised regression approaches. Recall the linear regression model can be expressed as:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{2.30}$$

where noise variables are independent and identically (iid) distributed normally distributed random variables:

$$\epsilon_i \sim N(0, \sigma^2), \quad i = 1, \ldots, n.$$

Without loss of generality, $\mathbf{y}$ are assumed to be standardised to have mean zero and $\mathbf{X}$ are assumed to be column-standardised to have mean zero and unit length.

Under a Bayesian framework, the Bayesian general linear model can be represented as [54]:

$$\mathbf{y}|\mathbf{X}, \boldsymbol{\beta}, \sigma^2 \sim \mathcal{N}_n \left(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I}_n\right). \tag{2.31}$$

The motivation of the Bayesian analysis is the fact that a good solution for $\boldsymbol{\beta}$ in the linear regression model (4.1) can be interpreted as the posterior mode of $\boldsymbol{\beta}$ from the Bayesian model when $\boldsymbol{\beta}$ follows a certain prior distribution. There are two main prior distributions: the spike and slab prior [55] and the continuous prior.

2.4.1 Spike and Slab Priors

A common way of estimating β in the Bayesian approaches is generating sparsity through a *spike and slab* prior. The idea was originally proposed by [55], and then further developed by [56].

A spike-and-slab prior is a mixture of two Gaussian distributions where one reaches a high point with low variance (the spike) and the other is broad with high variance (the slab). One widely used spike and slab prior for β, with the spike concentrating near zero and the slab spreading away from zero, is [56]:

$$\beta_j|\gamma_j \sim (1 - \gamma_j)\mathcal{N}(0, \tau_j^2) + \gamma_j \mathcal{N}(0, c_j \tau_j^2), \quad j = 1, \ldots, p, \tag{2.32}$$

where γ_j are binary variables, τ_j^2 is small and $c_j > 0$ is large. When $\gamma_j = 1$, the prior distribution for β has larger variance and therefore more likely have larger posterior values. However, a fully Bayesian approach with a spike and slab mixture for each β requires exploration of a model space of size 2^p, which becomes computationally difficult especially when p is large.

2.4.2 Global-Local Continuous Priors

Since most of the conventional Bayesian variable selection methods based on spike and slab priors rely on MCMC algorithms, it is computationally inefficient especially for high dimensional data. In addition, posterior sampling often requires stochastic search over a arge space of complicated models. When marginal likelihood is not analytically tractable, it convergents very slowly [57]. The potential computational issue with discrete mixture priors have motivated the Bayesian approaches with continuous shrinkage priors.

An alternative way of estimating β in the Bayesian approaches is by specifying both the spike and slab as continuous distributions. The Bayesian regularisation methods via the continuous shrinkage priors can be written as scale mixtures, which leads to a simpler MCMC algorithm with no marginal likelihood need to be calculated. Unlike conventional Bayesian methods with spike and slab priors, Bayesian regularisation methods specify shrinkage priors and this enables variable selection and coefficient estimation simultaneously.

There are many Bayesian approaches with continuous shrinkage priors such as the Bayesian ridge [58], the Bayesian Lasso [59], the Bayesian adaptive Lasso [60],

the Bayesian elastic net [61], the Bayesian horseshoe [62], the Bayesian horseshoe+ [63], etc. In general the Bayesian hierarchical model for global-local shrinkage priors [12] can be written as:

$$\begin{aligned} \boldsymbol{\beta}|\sigma^2, \tau^2, \lambda_1, \ldots, \lambda_p &\sim \mathcal{N}(\mathbf{0}, \sigma^2\tau^2\mathbf{D}_{\boldsymbol{\lambda}}), \quad \mathbf{D}_{\boldsymbol{\lambda}} = \text{diag}(\lambda_1^2, \ldots, \lambda_p^2) \\ \lambda_j &\sim \pi(\lambda_j)d\lambda_j, \ j = 1, \ldots, p \\ \tau &\sim \pi(\tau)d\tau \\ \sigma^2 &\sim \pi(\sigma^2)d\sigma^2, \end{aligned} \tag{2.33}$$

where λ_j are called the local shrinkage parameters that control the shrinkage for each predictors and τ is the global shrinkage parameter that controls the overall shrinkage. The common priors for τ include an exponential distribution and a standard half Cauchy distribution. There are various choices for the prior distribution of σ^2, and one of the widely used priors are the improper prior:

$$\pi(\sigma^2) \propto \frac{1}{\sigma^2}.$$

2.4.2.1 Bayesian Ridge Regression

Suppose ϵ_i is i.i.d. normal errors with mean zero and known variance σ^2, and the prior distributions for β_j are the independent normal priors:

$$\beta_j \sim N\left(0, \frac{\sigma^2}{\tau}\right),$$

or in matrix form:

$$\boldsymbol{\beta} \sim N\left(\mathbf{0}, \frac{\sigma^2}{\tau}\mathbf{I}_p\right),$$

There is no local shrinkage parameter in the Bayesian ridge model which implies that λ_j are equal to one in (2.33). A large value of the global shrinkage parameter τ corresponds to a prior that is more tightly concentrated around zero, and hence leads to stronger shrinkage towards zero.

Ridge regression improves stability in high variance problems such as regression with collinearity or many predictors. By shrinking coefficients towards zero, it reduces the variance.

2.4.2.2 Bayesian Lasso

As shown in Section 2.3.4, the Lasso estimate is [9]:

$$\hat{\boldsymbol{\beta}}_{\text{Lasso}} = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\arg\min} \left\{ \sum_{i=1}^{n} \left(y_i - \mathbf{x}_i^\top \boldsymbol{\beta}\right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\},$$

and it can also be expressed as a Bayesian posterior mode estimate under the assumption that β_j follows a Laplace distribution. The Bayesian Lasso [59, 64] has been introduced with the conditional priors of $\boldsymbol{\beta}$ having the following form:

$$p(\boldsymbol{\beta}|\sigma^2) = \prod_{j=1}^{p} \left(\frac{\tau}{2\sqrt{\sigma^2}} \exp\left\{ \frac{-\tau|\beta_j|}{\sqrt{\sigma^2}} \right\} \right),$$

with an improper scale invariant prior for σ^2 expressed as:

$$\pi(\sigma^2) \propto \frac{1}{\sigma^2}.$$

Conditioning on σ^2 leads to a unimodal full posterior [59].

The Bayesian Lasso hierarchical representation of the full model:

$$\begin{aligned} \boldsymbol{\beta}|\sigma^2, \tau^2, \lambda_1, \ldots, \lambda_p &\sim \mathcal{N}(\mathbf{0}, \sigma^2\tau^2\mathbf{D}_{\boldsymbol{\lambda}}), \quad \mathbf{D}_{\boldsymbol{\lambda}} = \text{diag}(\lambda_1^2, \ldots, \lambda_p^2) \\ \lambda_j^2 &\sim \text{Exp}(1), \quad j = 1, \cdots, p. \end{aligned}$$

The least squares, ridge and Lasso estimates can also be viewed as Bayes posterior modes under uniform, normal and Laplace priors for β_j's respectively. The Laplace distribution puts more weight near zero and less mass on two tails, whereas the uniform distribution is flat and puts weight equally. The normal distribution lies

somewhere between the two distributions. This explains why the Lasso regression tends to shrink some of coefficients toward zeros. Another important fact is that when the tunning parameter λ approaches zero, the Laplace distribution approaches to a uniform distribution as the penalty term disappears.

2.4.2.3 Bayesian Horseshoe

A competitive Bayesian alternative to the Lasso is the Bayesian model resulting from the horseshoe prior, which is a one-component prior [65]. The horseshoe arises from the same class of multivariate scale mixtures of normals as the Lasso, but it is almost always superior to the double-exponential prior at handling sparsity [62]. Furthermore, the horseshoe prior is known for its robustness at handling large outlying signals and the estimator of the horseshoe model does not face the computational issues of the point mass mixture models.

The horseshoe prior assumes that β_j are conditionally independent and each has a density function that can be written as a Gaussian scale mixtures. The Bayesian horseshoe prior is defined:

$$\begin{aligned} \boldsymbol{\beta}|\sigma^2, \tau^2, \lambda_1, \ldots, \lambda_p \sim \mathcal{N}(\mathbf{0}, \sigma^2\tau^2\mathbf{D}_{\boldsymbol{\lambda}}), \quad \mathbf{D}_{\boldsymbol{\lambda}} = \text{diag}(\lambda_1^2, \ldots, \lambda_p^2) \\ \lambda_j \sim C^+(0,1), \ j = 1, \ldots, p \\ \tau \sim C^+(0,1) \end{aligned} \tag{2.34}$$

where $C^+(0,1)$ is a standard half-Cauchy distribution with the probability density function:

$$f(x) = \frac{2}{\pi(1+x^2)}, \ x > 0. \tag{2.35}$$

The scale parameters λ_j are local shrinkage parameters and τ is the global shrinkage parameter.

The horseshoe prior leaves strong signals unshrunk and penalises noise variables severely. Therefore, the horseshoe prior has the ability to adapt to different sparsity patterns, while simultaneously avoiding over-shrinkage of large coefficients.

2.4.2.4 Bayesian Horseshoe+

To handle the ultra-sparse problem, the horseshoe prior can be extended to incorporate an extra level of hyper-parameters and the horseshoe+ prior is proposed [63]. The horseshoe+ prior therefore is an extension to the existing horseshoe prior and the Bayesian horseshoe+ model is defined as follows:

$$\begin{aligned} \boldsymbol{\beta}|\sigma^2, \tau^2, \lambda_1, \ldots, \lambda_p &\sim \mathcal{N}(\mathbf{0}, \sigma^2\tau^2\mathbf{D}_{\boldsymbol{\lambda}}), \quad \mathbf{D}_{\boldsymbol{\lambda}} = \text{diag}(\lambda_1^2, \ldots, \lambda_p^2) \\ \lambda_j|\phi_j &\sim C^+(0, \phi_j), \ j = 1, \ldots, p \\ \phi_j &\sim C^+(0, 1) \\ \tau &\sim C^+(0, 1) \end{aligned} \tag{2.36}$$

The horseshoe+ prior has a heavier tail than the regular horseshoe prior which enables larger success in separating the signals, and it has more aggressive shrinkage at the origin which leads to better handling of sparsity. Hence, the horseshoe+ prior produces sparser estimates than the regular horseshoe prior.

2.4.2.5 The Dirichlet-Laplace Prior

The Dirichlet-Laplace prior [13] for $\boldsymbol{\beta}$ denoted as $\boldsymbol{\beta} \sim \text{DL}_a$ is:

$$\begin{aligned} \beta_j|\sigma, \tau, \lambda_j &\sim \text{DE}(\sigma\tau\lambda_j) \\ (\lambda_1, \ldots, \lambda_p) &\sim \text{Dir}(a, \ldots, a) \\ \tau &\sim \text{Gamma}\left(pa, \frac{1}{2}\right), \end{aligned}$$

where the λ_j are the local scales, τ is the global scale, $\text{DE}(\cdot)$ denotes a zero mean double exponential or Laplace distribution and $\text{Dir}(a, \ldots, a)$ denotes a Dirichlet distribution with concentration parameter $(a, \ldots, a)$. The Dirichlet-Laplace prior

can be alternatively represented as:

$$\begin{aligned}\beta_j|\sigma^2, \psi_j, \lambda_j, \tau &\sim N(0, \sigma^2\tau^2\lambda_j\psi_j)\\ \psi_j &\sim \text{Exp}\left(\frac{1}{2}\right)\\ (\lambda_1, \ldots, \lambda_p) &\sim \text{Dir}(a, \ldots, a)\\ \tau &\sim \text{Gamma}\left(pa, \frac{1}{2}\right).\end{aligned}$$

The small values of a lead to most of λ_j close to zero and only few of them nonzero; while large values of a allow less singularity at zero, thus controlling the sparsity of regression coefficients.

The Dirichlet-Laplace prior leads to efficient posterior computation. Both simulations and real data examples has shown the Dirichlet-Laplace prior has good finite sample performance.

2.4.2.6 The R2-D2 Prior

Recently, a R-squared induced Dirichlet decomposition (R2-D2) prior [66] has been proposed. Unlike most other methods that put a prior on $\boldsymbol{\beta}$ directly, the R2-D2 prior is induced by putting a prior on a univariate function of $\boldsymbol{\beta}$, and then inducing a prior on the p-dimensional $\boldsymbol{\beta}$.

By assuming a kernel on each dimension of $\boldsymbol{\beta}$, the R2-D2 prior is proposed as:

$$\begin{aligned}\beta_j|\lambda_j, \tau &\sim K(\sigma^2\lambda_j\tau)\\ \lambda &\sim \text{Dir}(a_\pi, \ldots, a_\pi)\\ \tau &\sim \text{BP}(a, b),\end{aligned}$$

where $K(\delta)$ is a kernel with mean zero and variance δ, $\text{Dir}(a_\pi, \ldots, a_\pi)$ denotes a Dirichlet prior with concentration parameter $(a_\pi, \ldots, a_\pi)$ and $\text{BP}(a, b)$ is a Beta Prime distribution. Therefore, the R2-D2 prior is induced by a Beta prior on the coefficient of determination R^2, and then the total prior variance of the regression coefficients $\boldsymbol{\beta}$ is decomposed through a Dirichlet prior.

To encourage shrinkage, it is reasonable to assume that the kernel density $K(\delta)$ is a double exponential kernel, and therefore, the R2-D2 prior can be expressed as:

$$\begin{aligned}
\beta_j|\sigma^2, \psi_j, \lambda_j, \tau &\sim N\left(0, \frac{1}{2}\sigma^2\tau\lambda_j\psi_j\right) \\
\psi_j &\sim \text{Exp}\left(\frac{1}{2}\right) \\
(\lambda_1, \ldots, \lambda_p) &\sim \text{Dir}(a_\pi, \ldots, a_\pi) \\
\tau|\xi &\sim \text{Gamma}\,(a, \xi) \\
\xi &\sim \text{Gamma}(b, 1)
\end{aligned}$$

The R2-D2 prior has been proven to yield a consistent posterior. Compared to other shrinkage priors such as the horseshoe, the horseshoe+, the Dirichlet-Laplace, the R2-D2 prior possesses an unbounded density around zero with polynomial order, and the heaviest tails among these common shrinkage priors.

One desirable property of the prior distributions is that it has a heavy tail because it allows important predictors (signal) to be identified for the high-dimensional regression. Compared to horseshoe, horseshoe+ and Dirichlet-Laplace model, the R2-D2 has been shown to lead to the heaviest tail, followed by the horseshoe+, then the horseshoe, and finally the DL prior.

Another desirable property is that the prior distributions have a more mass around zero so that most predictors have little or no effect (noise) on the response variable. The R2-D2 prior and DL prior are proven to be more efficient (i.e. higher concentration around zero) than the horseshoe+, and then followed by the horseshoe+.

2.4.2.7 Posterior Sampling

To compute the posterior samples from the Bayesian model (2.33), the Gibbs sampler is implemented (see Section 2.1.7.2). The Gibbs sampler cyclically updates one parameter at a time given the current values of all the other parameters and hyper-parameters.

The posterior distributions for $\boldsymbol{\beta}$ and σ^2 do not depend on λ_j. Therefore, the full conditional distributions of $\boldsymbol{\beta}$ and σ^2 can be derived from the joint posterior distribution using the Gibbs sampler, and are the same for all the models introduced above. The full conditional distribution of $\boldsymbol{\beta}$ is

$$\begin{aligned}\boldsymbol{\beta} \,|\cdot &\sim \mathcal{N}\left(\mathbf{A}^{-1}\mathbf{X}^T\mathbf{y}, \sigma^2\mathbf{A}^{-1}\right)\\ \mathbf{A} &= \mathbf{X}^T\mathbf{X} + (\tau^2\mathbf{D}_{\boldsymbol{\lambda}})^{-1}.\end{aligned}$$

The full conditional distribution of σ^2 is

$$\sigma^2 \,|\cdot \sim \mathcal{IG}\left(\frac{n-1+p}{2}, \frac{(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^T(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) + \boldsymbol{\beta}^T(\tau^2\mathbf{D}_{\boldsymbol{\lambda}})^{-1}\boldsymbol{\beta}}{2}\right),$$

where $\mathcal{IG}$ is the inverse Gamma distribution.

One concern of the horseshoe model is that the conditional posterior distribution for the hyper-parameter is difficult to sample from using the Gibbs sampler due to the non-standard form caused by the half-Cauchy distribution. This issue was addressed in [67] and enables a straightforward sampling of the full conditional posterior distributions by decomposing

$$\tau \sim C^+(0, a) \tag{2.37}$$

into

$$\tau^2 \,|\nu \sim \mathcal{IG}\left(\frac{1}{2}, \frac{1}{\nu}\right) \text{ and } \nu \sim \mathcal{IG}\left(\frac{1}{2}, \frac{1}{a^2}\right). \tag{2.38}$$

2.4.2.8 Sparsifying Posterior

While continuous global-local shrinkage priors exhibit desirable theoretical, computational and empirical properties, they also have their own challenges. Unlike the discrete mixture priors which directly generate sparse estimates, shrinkage priors require additional steps to go from the continuous posterior distribution to a sparse estimate.

Thresholding Rule There are several methods to deal with this. The most common method is to use a threshold to decide which predictor is to be included. For example, [62] described a simple thresholding rule for the horseshoe prior that yields a sparse estimate. Similar to the Bayesian discrete-mixture model, the horseshoe estimator gives weights, and we call the predictor a signal if the weight is greater than or equal to 0.5 and a noise if the weight is less than 0.5.

In addition to thresholding, there are two major approaches: penalised variable selection based on posterior credible regions, and decoupling shrinkage and selection. We will now introduce these two methods.

Penalised Variable Selection Based On Posterior Credible Regions Under the setting of large p and small n, Bayesian variable selection via penalised credible regions [68] has been proposed. This method fits the full model with all predictors included using a continuous shrinkage prior, constructs a series of joint credible regions based on the posterior distribution, and then selects the sparsest solution within each posterior credible region.

Formally, the prior distribution is chosen as

$$\begin{aligned} \boldsymbol{\beta}|\sigma^2, \tau &\sim N\left(0, \frac{\sigma^2}{\tau}\mathbf{I}_p\right) \\ \sigma^2 &\sim \text{InvGamma}(0.01, 0.01) \end{aligned}$$

where τ can be either fixed or given a Gamma distribution.

From this model, the $(1-\alpha)100\%$ posterior credible regions C_α is an elliptical credible region that can be computed as:

$$C_\alpha = \left\{\boldsymbol{\beta} : (\boldsymbol{\beta} - \bar{\boldsymbol{\beta}})^\top \bar{\boldsymbol{\Sigma}}^{-1} (\boldsymbol{\beta} - \bar{\boldsymbol{\beta}}) \leq C_\alpha\right\},$$

where $\bar{\boldsymbol{\beta}}$ and $\bar{\boldsymbol{\Sigma}}$ are posterior mean and covariance. The credible regions is to find $\tilde{\boldsymbol{\beta}}$ such that:

$$\tilde{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \|\boldsymbol{\beta}\|_0 \quad \text{s.t.} \quad \boldsymbol{\beta} \in C_\alpha,$$

where $\|\boldsymbol{\beta}\|_0$ is the number of nonzeros in $\boldsymbol{\beta}$. The optimisation problem can be alternatively expressed by replacing ℓ_0 with a smooth bridge between ℓ_0 and ℓ_1 and by using a linear approximation:

$$\tilde{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} (\boldsymbol{\beta} - \bar{\boldsymbol{\beta}})^\top \bar{\boldsymbol{\Sigma}}^{-1} (\boldsymbol{\beta} - \bar{\boldsymbol{\beta}}) + \lambda_\alpha \sum_{j=1}^{p} \frac{\beta_j}{|\bar{\beta}_j|^2}.$$

Since there is one-to-one correspondence between λ_α and α, the LARS algorithm can be used for solving the full path for various values of α. This proposed method builds connections between Bayesian and penalised regression methods.

Decoupling Shrinkage and Selection Method The other sparsifying method proposed to find sparse sets of variables is the decoupling shrinkage and selection method (DSS) [14]. The DSS uses a generic loss function combining a posterior summariser with an explicit parsimony penalty to induce sparse estimator.

In the parameter estimation problem, a loss function L is normally used to quantify the difference between estimated and true values of the data. The"decoupled shrinkage and selection" (DSS) loss function is proposed as:

$$L(\boldsymbol{\gamma}) = \lambda \|\boldsymbol{\gamma}\|_0 + n^{-1} \|\mathbf{X}\bar{\boldsymbol{\beta}} - \mathbf{X}\boldsymbol{\gamma}\|_2^2, \tag{2.39}$$

where $\|\cdot\|_0 = \sum_j \mathbb{1}(\gamma_j \neq 0)$, $\bar{\boldsymbol{\beta}} = \mathrm{E}(\boldsymbol{\beta}|\mathbf{y})$ is the posterior mean of samples.The optimised solution of the DSS loss function then represents a sparsification of $\bar{\beta}$, and has the following form:

$$\boldsymbol{\beta}_\lambda = \underset{\boldsymbol{\gamma}}{\operatorname{argmin}}\ \lambda \|\boldsymbol{\gamma}\|_0 + n^{-1} \|\mathbf{X}\bar{\boldsymbol{\beta}} - \mathbf{X}\boldsymbol{\gamma}\|_2^2. \tag{2.40}$$

The optimal solution is achieved through the trade-off between the number of variables in the model and the predictive performance in terms of the mean squared prediction error. The best sparse estimates that approximate the posterior mean of the samples are computed through the correlation of X, and thus, correlation in the data is also taken into account. If a variable X_j has strong association

with the response variable, the posterior mean $\bar{\beta}_j$ of this variable will tend to be large, and therefore, X_j is more likely to be included in the final sparsified model. Conversely, if a variable X_j has no effect or little effect on the response variable, it is more likely to be removed from the final model as $\bar{\beta}_j$ is small. Therefore, the DSS algorithm builds connections between Bayesian-type methods to variable selection and popular penalised likelihood approaches.

2.4.3 Example

In this section, we illustrate the performance of Bayesian regression based on different global-local continuous priors using the diabetes data. There are $n = 442$ patients and $p = 10$ predictors for the dataset. The predictors are age, sex, body mass index (BMI), average blood pressure (BP) and six blood serum measurements (S1, S2, S3, S4, S5, S6). The response variable is a quantitative measure of disease progression one year after baseline.

Before performing analysis on diabetes data, we standardised the response variable to have a zero mean and the covariates to have a zero mean and unit variance:

$$\sum_{i=1}^{n} y_i = 0, \ \sum_{i=1}^{n} x_{ij} = 0, \ \sum_{i=1}^{n} x_{ij}^2 = 1, \quad \text{for } j = 1, \cdots, p.$$

Three methods were implemented on the diabetes data, non-negative garrote discussed in Section 2.3.3, Lasso in Section 2.3.4 with 10-fold cross-validation and least angle regression in Section 2.3.4.1. The solution paths for the above three methods are shown in Figure 2.5. The NNG analysis of the diabetes data shown on the left panel of Figure 2.5 indicates all NNG solutions for $\hat{\boldsymbol{\beta}}$ as $\sum |\beta_j|$ increases from 0 to 3459.98. When $\sum |\beta_j|$ reaches 3459.98, the estimate is simply the least squares estimate. The order of the variables enter the NNG model is: X_9, X_3, X_4, X_5, X_2, X_8, X_6, X_{10}, X_7, X_1. The middle panel shows the solution paths of Lasso with 10-fold cross-validation, and indicates the order of variables enter the model is: variable X_3 and X_9, X_4, X_7, X_2, X_{10}, X_5, X_8, X_6, X_1. Here, variables X_3 and X_9 enter the model simultaneously. The right panel is the solution path of LARS

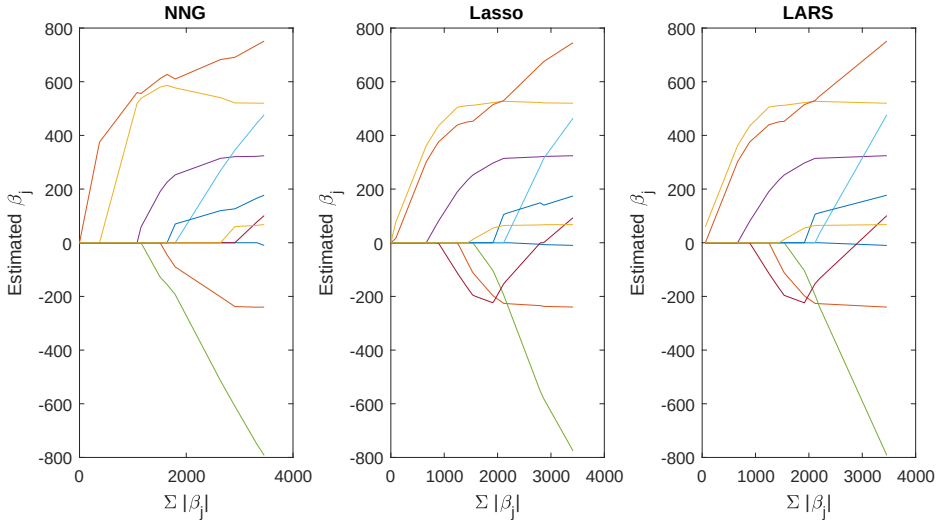

FIGURE 2.5: Plots of regression coefficients $\hat{\beta}_j$ versus sum of absolute values of estimates $\sum|\hat{\beta}_j|$ for non-negative garrote, Lasso with cross-validation method and least angle regression.

where it is almost identical to the solution path of Lasso with 10-fold CV, and the only difference is that variable X_3 enters the model before variable X_9.

We also ran the Bayesian model with ridge, Lasso, horseshoe and horseshoe+ priors. For each prior distribution, we ran the Bayesian algorithm using the 'BayesReg' software toolbox [69] for the sampling procedure, discarded first $10,000$ samples as burn-in samples, and obtained $100,000$ posterior samples. The software can be downloaded from MathWorks File Exchange 60823. The mean of posterior samples was calculated as our estimates. The estimates of standardised coefficients for Bayesian methods using different priors, along with the least squares estimate are shown in Table 2.1. The ratio of $\sum_{j=1}^{p}|\hat{\beta}_j|/\sum_{j=1}^{p}|\hat{\beta}_j^{\mathrm{LS}}|$ was also computed and shown on the last line of the table.

The estimate of Lasso CV is quite close to the estimate of Bayesian Lasso estimate. However, the Bayesian approaches do not yield sparse estimates. To produce a sparse model using posterior means, Tang et al. [70] proposed a criterion for selecting the active predictors by computing the ratio of $|\hat{\beta}_j|/|\hat{\beta}_j^{\mathrm{LS}}|$. If the ratio is greater than 0.5, the predictor is selected as active variables and can be added to the final model. The inactive predictors are indicated by asterisk in Table 2.1. From the table, it is noted that both ridge and Lasso models treat variables X_1,

X_5 and X_6 as inactive predictors, the horseshoe has added one more variable X_8 as inactive predictors, and horseshoe+ model has further removed variable X_{10} from the model. Compared to the ridge and Lasso priors, the horseshoe prior leads to a sparse final model and the horseshoe+ prior produces a sparser result. The overall ratio for the ridge model gives the largest value, followed by the Lasso model, the horseshoe, and the horsehoe+ produces the smallest ratio.

Variable	Least squares	Lasso CV	Bayesian ridge	Bayesian Lasso	Bayesian hs	Bayesian hs+
Age	-10.01	0.00*	-3.89*	-3.62*	-2.62*	-1.86*
SEX	-239.82	-219.12	-224.98	-211.40	-196.86	-193.31
BMI	519.85	525.74	511.55	523.28	535.24	538.41
BP	324.38	310.12	313.99	306.03	301.53	303.00
S1	-792.18	-172.53*	-188.33*	-183.47*	-165.86*	-154.93*
S2	476.74	0.00*	1.44*	5.89*	7.85*	10.56*
S3	101.04	-170.19	-155.56	-152.51	-156.94	-164.13
S4	177.06	80.00*	116.15	97.51	70.98*	57.86*
S5	751.27	526.16	507.23	522.43	536.08	539.35
S6	67.63	62.09	76.97	64.13	42.75	30.25*
$\frac{\|\hat{\boldsymbol{\beta}}\|_1}{\|\hat{\boldsymbol{\beta}}_{\mathrm{LS}}\|_1}$	1.00	0.597	0.607	0.598	0.583	0.576

TABLE 2.1: Least squares estimate and mean estimates of regression coefficients of Bayesian ridge, Lasso, horseshoe (hs), horseshoe+ (hs+) models for the diabetes data.

2.5 Group Structures

Group structures naturally exist in many statistical applications. Often in the regression problems, variable selection aims to select groups of variables rather than individual variables. For example, a multi-level categorical predictor in the regression model can be represented by a group of dummy variables, a continuous predictor in the additive model can be expressed as a group of basis functions, and prior knowledge such as genes in the same biological pathway form a natural group.

Under the assumption that the group structure is given, the aim is to find sparsity at both the group level, as well as within the groups themselves. Suppose there are total G number of groups in the linear regression model, then the regression

model can be expressed as:

$$\mathbf{Y} = \sum_{g=1}^{G} \mathbf{X}_g \boldsymbol{\beta}_g + \boldsymbol{\epsilon}, \tag{2.41}$$

where $\mathbf{X}_g$ is the submatrix of $\mathbf{X}$ containing all predictors in group g, $\boldsymbol{\beta}_g$ are the regression coefficients associated with group g.

2.5.1 Non-Bayesian Approaches

To perform group-wise selection, many penalised linear regression methods have been proposed for grouped variables. For example, the group Lasso [10], the group non-negative garrotte [71], the group adaptive Lasso [72]. In the following sections, we will introduce those methods.

2.5.1.1 Group Lasso

One of the important group selection procedures is the group Lasso [37], which extends the Lasso to perform group-wise selection. For positive definite matrices $K_1, \ldots, K_J$, the group Lasso estimate is defined as the solution to the convex optimisation problem:

$$\frac{1}{2}\|\mathbf{y} - \sum_{g=1}^{G} \mathbf{X}_g \boldsymbol{\beta}_g\|_2^2 + \lambda \sum_{g=1}^{G} \|\boldsymbol{\beta}_g\|_{\mathbf{K}_g},$$

where $\|\boldsymbol{\beta}_g\|_{\mathbf{K}_g} = (\boldsymbol{\beta}_g' \mathbf{K}_g \boldsymbol{\beta}_g)^{1/2}$ and $\mathbf{K}_g$ are symmetric positive definite matrices. However, a difficulty in using the group Lasso is that the corresponding solution path is no longer piecewise linear and is potentially slow to compute.

2.5.1.2 Group LARS

The group version of the LARS algorithm is an extension of the LARS and it is defined as follows:

- Initialise $\boldsymbol{\beta}^{(0)} = 0,\ k = 1,\ \mathbf{r}^{(0)} = Y$;
- Compute the current most correlated set $\mathcal{A}_1 = \arg\min_g ||\mathbf{X}_g' \mathbf{r}^{(k-1)}||^2 / s_g$;
- Compute current direction $\gamma_{\mathcal{A}_k} = (\mathbf{X}'_{\mathcal{A}_k} \mathbf{X}_{\mathcal{A}_k})^{-1} \mathbf{X}'_{\mathcal{A}_k} \mathbf{r}^{(k-1)}$ where $\gamma_{\mathcal{A}_\gamma^c} = 0$;
- For every $g \notin \mathcal{A}_k$, compute $\|\mathbf{X}_g'(\mathbf{r}^{(k-1)} - \alpha_g \mathbf{X}_\gamma)^2\|/s_g = ||\mathbf{X}_g'(\mathbf{r}^{(k-1)} - \alpha_j \mathbf{X}_\gamma)^2||/s_g'$, where $\alpha_j \in [0, 1]$;
- If $\mathcal{A}_k \neq \{1, \ldots, G\}$, let $\alpha = \min_{g \notin \mathcal{A}_k}(\alpha_g) = \alpha_g$ and update $\mathcal{A}_{k+1} = A \cup \{g^*\}$; otherwise set $\alpha = 1$;
- Update $\boldsymbol{\beta}^{(k)} = \beta^{(k-1)} + \alpha\gamma, \mathbf{r}^{(k)} = Y - \mathbf{X}\boldsymbol{\beta}^{(k)}$ and $k = k + 1$
- Repeat the above iterations until $\alpha = 1$.

The computational advantage of the group LARS algorithm is that it is piecewise linear and therefore group LARS has comparable performance with the group Lasso. Whereas the group Lasso does not possess this piecewise linear property unless each group has only one predictor or unless X is orthonormal. As a result, the group Lasso is computationally more expensive in high-dimensional problems then group LARS. The solution paths for the group Lasso and group LARS are usually similar as long as $\max_g(s_g)$ is not very large. Both approaches give excellent performance in selecting grouped variables [37].

2.5.1.3 Group Non-Negative Garrotte

Another popular grouped model is the group non-negative garrotte (NNG) [37] whose formulation is similar to the adaptive Lasso. The group NNG reduces the dimensionality of the solution path to the number of groups, G. An approximate solution can be achieved by solving:

$$\mathbf{d}_\lambda = \underset{\mathbf{d}}{\operatorname{argmin}} \left(\frac{1}{2} \|\mathbf{y} - \mathbf{Z}\mathbf{d}\|^2 + \lambda \sum_{g=1}^{G} s_g d_g \right), \quad \text{subject to } d_g \geq 0, \forall g, \tag{2.42}$$

where $\mathbf{Z} = (\mathbf{Z}_1, \cdots, \mathbf{Z}_G)$, $\mathbf{Z}_g = \mathbf{X}_g\bar{\boldsymbol{\beta}}_g$, s_g is the size of group g and $\mathbf{d} = (d_1, \ldots, d_G)$ is a vector of group shrinkage coefficients. The algorithm for computing the non-negative garotte solution path is similar to the group least angle regression algorithm, and the solution for a given λ can be computed as $\boldsymbol{\beta}_\lambda = (\bar{\boldsymbol{\beta}}_1 d_1, \cdots, \bar{\boldsymbol{\beta}}_G d_G)$.

2.5.1.4 Adaptive Group Lasso

Although the group Lasso respecting the grouping structure in the data makes the group Lasso attractive, it is in general not selection consistent [37] and it is inclined to select more groups than what are in the model.

Therefore, Wei and Huang [73] proposed an adaptive group Lasso method which is a generalisation of the adaptive Lasso [11] and it requires an initial estimator.

To generalise the adaptive Lasso for individual variable selection to group selection, consider a general group Lasso with a weighted penalty term:

$$\frac{1}{2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + \tilde{\lambda}\sum_{g=1}^{G} w_g\sqrt{d_g}||\boldsymbol{\beta}_g||_2,$$

where w_g is the weight associated with the gth group and tunning parameter for the gth group in the general form λ_g can be regarded as $\tilde{\lambda}w_g$ in this equation. For groups with large coefficients, the penalty can be small while for groups with small coefficients, the penalty can be large. To gain the information about the coefficients, a consistent initial estimator can be applied.

A simple way of determining the weight w_g is to use the initial estimator say $\tilde{\boldsymbol{\beta}}$. Then w_g can be defined as:

$$w_g = \frac{1}{||\tilde{\boldsymbol{\beta}}_g||_2}, \ g = 1, \ldots, G.$$

Recall the adaptive Lasso estimator is

$$\hat{\boldsymbol{\beta}}^{*(n)} = \arg\min_{\boldsymbol{\beta}} ||\mathbf{y} - \sum_{g=1}^{G} \mathbf{x}_g\boldsymbol{\beta}_g||^2 + \lambda_n \sum_{g=1}^{G} \hat{w}_g|\boldsymbol{\beta}_g|,$$

where $\hat{\mathbf{w}} = 1/|\hat{\boldsymbol{\beta}}|^{\gamma}$. Therefore, this choice is of the penalty levels for each group is a natural generalisation of the adaptive Lasso. In particular, if there is one element in each group, the weight is exactly same as the adaptive Lasso penalty.

It is shown that the adaptive group Lasso is consistent in group selection under certain conditions provided that an initial estimator is estimation consistent and does not miss important groups. Moreover, the convergence rate of the adaptive group Lasso is improved compared with that of the group Lasso if an appropriate penalty parameter $\tilde{\lambda}$ is selected. In summary, the iterated group Lasso procedure, which uses the group Lasso to obtain an initial estimator for dimensionality reduction, and then uses the adaptive group Lasso to reach selection consistency, is an attractive approach to group selection in high-dimensional settings.

2.5.2 Bayesian Approaches

Alternative to the non-Bayesian approaches to group-wise selection, Bayesian approaches have also been proposed.

2.5.2.1 Bayesian Group Lasso

Continuous Priors Raman et al. [74] proposed the Bayesian version of the Group-Lasso which extend the standard Lasso by using a hierarchical expansion to handle the grouped variables.

A hierarchical Normal-Gamma expansion of the Multi-Laplace prior on the regression coefficients is derived to obtain an efficient Gibbs sampling for the Group-Lasso model. For grouped variables, the prior distributions of $\boldsymbol{\beta}_g$ can be expressed as a two-level hierarchical model. Define $a_g = s_g\rho$ and $b_g = ||\boldsymbol{\beta}_g||_2^2/\sigma^2$ for each group g, then the Multi-Laplacian prior on $\boldsymbol{\beta}_g$ is hierarchical Normal-Gamma

model:

$$p(\boldsymbol{\beta}_g|\rho) = \int_0^\infty N(\boldsymbol{\beta}_g|0, \sigma^2\delta_g^2)\text{Gamma}\left(\delta_g^2|\frac{s_g+1}{2}, \frac{2}{a_g}\right)d\delta_g^2$$
$$\propto \text{M-Laplace}\left(\boldsymbol{\beta}_g|\mathbf{0}, \left(\frac{a_g}{\sigma^2}\right)^{-1/2}\right)$$

The Bayesian hierarchy model of group Lasso is:

$$\begin{aligned}\mathbf{y}|\mathbf{X}, \boldsymbol{\beta}, \sigma^2 &\sim \text{N}_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n),\\ \boldsymbol{\beta}_g &\sim \text{N}\left(0, \left(\frac{a_g}{\sigma^2}\right)^{-1/2}\right)\\ \delta_g^2 &\sim \text{Gamma}\left(\frac{s_g+1}{2}, \frac{2}{a_g}\right),\ g = 1, \cdots, G\end{aligned}$$

Therefore, the full conditional posteriors of the Group-Lasso with a Gaussian likelihood, Gaussian priors over the regression coefficients and Gamma priors over the corresponding variances are easily computed using the Gibbs sampler.

The Bayesian group Lasso [75] hierarchical model can also be expressed:

$$\begin{aligned}\mathbf{y}|\mathbf{X}, \boldsymbol{\beta}, \sigma^2 &\sim \text{N}_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n),\\ \boldsymbol{\beta}_g|\sigma^2, \delta_g^2 &\sim \text{N}_{s_g}(0, \sigma^2\delta_g^2\mathbf{I}_{s_g})\\ \delta_g^2 &\sim \text{Gamma}\left(\frac{s_g+1}{2}, \frac{\lambda^2}{2}\right),\ g = 1, \cdots, G\\ \sigma^2 &\sim \frac{1}{\sigma^2}d\sigma^2,\end{aligned}$$

where s_g is the number of predictors in group g and δ_g is shrinkage parameter for grouped variables.

Spike and Slab Priors Xu and Ghosh [76] first proposed the Bayesian group Lasso with spike and slab priors (BGL-SS) which consists of a multivariate point mass mixture prior and produces exact zero estimates at the group level to perform

variable selection. The hierarchical Bayesian models are:

$$\begin{aligned}
\mathbf{y}|\mathbf{X},\boldsymbol{\beta},\sigma^2 &\sim \mathrm{N}_n(\mathbf{X}\boldsymbol{\beta},\sigma^2\mathbf{I}_n),\\
\boldsymbol{\beta}_g|\sigma^2,\delta_g^2 &\sim (1-\pi_0)\mathrm{N}_{s_g}(0,\sigma^2\delta_g^2\mathbf{I}_{s_g})+\pi_0 m_0(\boldsymbol{\beta}_g),\ g=1,\cdots,G,\\
\delta_g^2 &\sim \text{Gamma}\left(\frac{s_g+1}{2},\frac{\lambda^2}{2}\right),\ g=1,\cdots,G,\\
\sigma^2 &\sim \text{InvGamma}(\alpha,\gamma)
\end{aligned}$$

where $\pi_0 \sim \text{Beta}(a,b)$ and $m_0(\boldsymbol{\beta}_g)$ is a point mass at $\mathbf{0}$. The full posterior distribution of all unknown parameters conditional on the data is easy to compute. The marginal posterior median is shown to be a soft thresholding estimator and therefore can select variables automatically.

The BGL-SS for group level variable selection is not always suitable for the problem. For example, a SNP that associated with a disease does not necessarily imply that all other SNPs in the same gene are related to the disease.

Xu and Ghosh [76] then proposed a bi-level selection method for selecting variables both at group level and within groups. A full Bayesian hierarchical model of the sparse group Lasso (BSGL) is:

$$\begin{aligned}
\mathbf{Y}|\mathbf{X},\boldsymbol{\beta},\sigma^2 &\sim \mathrm{N}_n(\mathbf{X}\boldsymbol{\beta},\sigma^2\mathbf{I}_n),\\
\boldsymbol{\beta}_g|\boldsymbol{\tau}_g,\gamma_g,\sigma^2 &\sim \mathrm{N}_{s_g}(0,\sigma^2\mathbf{V}_g),\ g=1,\cdots,G,\\
\pi(\delta_{g1}^2,\cdots,\delta_{gs_g}^2,\gamma_g^2) &= c_g(\lambda_1^2,\lambda_2^2)\prod_{j=1}^{s_g}\left[(\delta_{gj}^2)^{-1/2}\left(\frac{1}{\delta_{gj}^2}+\frac{1}{\gamma_g^2}\right)^{-1/2}\right](\gamma_g^2)^{-1/2}\\
&\quad\times\exp\left\{-\frac{\lambda_1^2}{2}\sum_{j=1}^{s_g}\delta_{gj}^2-\frac{\lambda_2^2}{2}\gamma_g^2\right\}.
\end{aligned}$$

With above hierarchical priors, the marginal prior on $\boldsymbol{\beta}_g$ is :

$$\pi(\boldsymbol{\beta}_g|\sigma^2)\propto\exp\left\{-\frac{\lambda_1}{\sigma}||\boldsymbol{\beta}_g||_1-\frac{\lambda_2}{\sigma}||\boldsymbol{\beta}_g||_2\right\}.$$

The BSGL has shrinkage effects at both the group level and within a group, but it does not produce exact zero estimators and therefore never produce sparse model.

The Bayesian sparse group selection with spike and slab prior (BSGS-SS) [77] therefore has been proposed to produce sparse solutions at group level as well as within groups by using spike and slab type priors for both group variable selection and individual variable selection. The coefficients $\boldsymbol{\beta}_g$ is re-parameterised to handle two kinds of sparsity separately:

$$\boldsymbol{\beta}_g = \mathbf{V}_g^{-1/2}\mathbf{b}_g,$$

where $\mathbf{V}_g = \text{diag}\{\delta_{g1}, \cdots, \delta_{gs_g}\},\ \delta_{gj} \geq 0,\ g = 1, \cdots, G,\ j = 1, \cdots, s_g$, and $\boldsymbol{\beta}_g$ is multivariate normal distribution with mean 0 and identity matrix as covariance matrix. To select variables at group level, the spike and slab prior for each $\mathbf{b}_g$ is:

$$\mathbf{b}_g \sim (1-\pi_0)\text{N}_{mg}(\mathbf{0}, \mathbf{I}_{s_g}) + \pi_0 m_0(\mathbf{b}_g),\ g = 1, \cdots, G.$$

To select variables within each group, the spike and slab prior for each δ_{gj} is:

$$\delta_{gj} \sim (1-\pi_1)\text{N}^+(0, s^2) + \pi_1 m_0(\delta_{gj}),\ g = 1, \cdots, G,\ j = 1, \cdots, s_g,$$

where $\text{N}^+(0, s^2)$ is a normal distribution truncated below at 0. Similar conditional posterior distributions can be derived.

2.6 Summary

In this section, we have introduced the Bayesian inference, Bayes' theorem, linear regression model, some non-Bayesian penalised linear regression models such as the ridge, the Lasso, the horseshoe, the horseshoe+. We have also reviewed information criteria, Bayesian penalised regreesion, and Bayesian linear regression with sparsifying methods such as the penalised variable selection via credible regions and decoupled shrinkage and selection methods. After that, we have introduced group structures, and existing literature of grouped selection methods.

Chapter 3

Genome-Wide Association Studies

In this chapter, genome-wide association studies (GWAS) are introduced. Then it is followed by some basic genetics topics. The conventional and modern approaches for GWAS are aslo included in this chapter.

3.1 Genome-Wide Association Studies (GWAS)

In this section, we will start with information about human genetics, and then formally introduce GWAS. After that, we introduce the conventional approaches to GWAS, and some most recent approaches to GWAS. Finally, we introduce breast cancer along with its risk predictors and impact, then review current GWAS approaches and findings for breast cancer.

3.1.1 General Definitions and Concepts of Genetics

Humans have units at a molecular level called genes which are inherited from their parents. In fact, all living organisms have genes that are contiguous stretches of

deoxyribonucleic acid (DNA). The human genome consists of basic biological material that can be inherited from parents to their children. This heritable material is stored in chromosomes in the nucleus of every cell. There are 23 pairs of chromosomes, of which 22 pairs are autosomes and the 23rd pair is the sex chromosome. Each chromosome consists of long strands of DNA which is composed of the pairs of four bases: A,C,T,G [78].

A genetic locus is a location in a chromosome that is polymorphic which implies that data can have more than one possible variant. The different variants at a locus are called alleles. Suppose we label two possible variants as A and a. For the autosomes, an individual with AA or aa is called homozygous; an individual with Aa is called heterozygous. Phenotypes are observable characteristics and thus a person's phenotype can be obtained without knowing their gene variant.

A genetic marker is a gene or a DNA sequence with a particular locus on a chromosome that can be used to identify individuals. Single Nucleotide Polymorphisms (SNPs) are the simplest types of genetic markers. Gene mapping involves any method used to find the chromosome location of a disease-causing gene and generally involves scanning the entire human genome of millions of genetic markers to look for genetic variation associated with disease traits. The term 'cases' refers to subjects with certain disease or condition in which researchers are interested while the term 'controls' refers to subjects that do not have the disease or condition.

3.1.2 Introduction to Genome-Wide Association Studies

Human genetics refers to the study of inheritance as it occurs in human beings. The genetic origin of diseases has been widely discussed for a long time. However, the discovery of the actual identification of the gene alleles for certain diseases such as schizophrenia, sickle cell anemia and cancers has been very limited. There are many different technologies including study designs and analytical tools for identifying genetic risk factors. Among those technologies, GWAS have evolved over the past decades and becomes a powerful tool for investigating genetic risk factors

of human disease. GWAS measures and analyses DNA sequence variations from the human genome and see if certain variant is associated with diseases. This is generally done by comparing the DNA of people with disease and without disease. The ultimate goal of GWAS is to improve the outcomes by using genetic risk factors in conjunction with developing preventive strategies, improve diagnostic tools and work on treatments [79–83].

To analyze DNA sequence variations of human genetics for GWAS, a basic unit of genetic variation is needed. SNPs are the simplest types of genetic markers and they contain two alleles which means that for one given SNP location, there are two base-pair possibilities in a population. The frequency of these alleles in the total population can be estimated from their frequency in a sample. One of those two alleles will appear less frequently than the other which is called the minor allele. The frequency of a SNP can be expressed as the minor allele frequency (MAF) which also refers to the frequency at which the least common allele occurs in a given population. Typically GWAS are designed to exclude SNPs with a MAF less than 5%.

SNPs have been extensively used as a way of investigating genetic causes of diseases. There are three main reasons why a SNP can be associated with a disease [84]:

1. The SNP itself is causative;
2. The SNP is in linkage disequilibrium with a close causative SNP;
3. The association is identified because of the population stratification.

There are roughly ten million common SNPs in the human genome, therefore scanning the entire genome of a large amount of patients would be time-consuming and expensive. Thanks to the International HapMap Project, those ten million variants are grouped into local neighborhoods, called ‘haplotypes’. This significant breakthrough has improved the efficiency and reduced the required number of

SNPs from ten million to 300,000. The HapMap project is a process that targets the SNPs with a minor allele frequency of 5% or greater.

In terms of the study design of GWAS, there are two types: family based studies and population based studies which also include case-control and cohort studies. Population-based studies are standard, practical ways of association analysis since they are studies of unrelated individuals and the data can be relatively easy to obtain while data is very difficult to gather for family based studies [85]. Therefore, they are more practical than family based studies [79, 86, 87].

However, a concern with population based studies is that these approaches are prone to population stratification, which can lead to biased or spurious results. To address the issue, various methods [88] such as principal components analysis can be applied [89].

3.1.2.1 Linkage Disequilibrium

Linkage disequilibrium (LD) refers to the linkage between markers in a population and it is another way of characterizing the degree of correlation between an allele of one SNP and an allele of another SNP. The term linkage refers to two markers on the same chromosome through generations. Recombinations from generation to generation will break the chromosomal segments, repeated random recombinations in a fixed-size population will break apart segments of contiguous chromosome and thus lead to linkage equilibrium in the population. The level of linkage disequilibrium is affected by a number of factors including population size, genetic linkage, selection, the rate of recombination, the rate of mutation, genetic drift, non-random mating, and population structure [90].

3.1.2.2 Model Set-Up

One of the important aspects in the study design is to characterize the phenotype of interest. There are two types of phenotypes: quantitative and categorical. Quantitative traits are quantitative measures of a disease, such as body mass index

(BMI), high-density lipoprotein cholesterol (HDL), intermediate-density lipoprotein (lDL). In general, the categorical phenotype is also known as binary or dichotomous and it generally refers to case/control. In this case, subjects are divided into two groups: affected and unaffected.

In current genome-wide association studies, the SNPs are tested one at a time for the association with the phenotype and therefore result in a series of single-locus statistics. The allelic association tests and genotypic association tests are two most common methods for identifying susceptibility variants. The allelic association test is a test that examines the relationship between one allele of the SNP and the phenotype. The genotypic association tests study the relationship between the genotypes and the phenotype. The genotypes for a SNP also vary from dominant, recessive, multiplicative and additive models.

In the statistical test, phenotypes and genotypes are indicated by the response variable y and regressors X. For dichotomous traits, the response variable y can be expressed as a vector consisting of 0's and 1's where 0 corresponds to a control and 1 corresponds to a case. For quantitative traits, y is a continuous number indicating the measure of interest. In terms of regressors X, there are several different ways of encoding the genotype data. If we assume two alleles of a SNP are 'A' and 'a', then a dominant model means that having one or more 'A' alleles increases the risk of getting disease. Hence, individuals with 'AA' or 'Aa' are compared with individuals with 'aa'. On the other hand, a recessive model makes the assumption that individuals with 'AA' are compared with individuals with 'Aa' or 'aa'. For an additive model, the number of minor alleles is generally counted and is coded as 2 for homozygous genotype of minor allele: 'AA', 1 for the heterozygous genotype 'Aa', and 0 for homozygous 'aa'. In the GWAS model, each genotype is generally encoded using additive model.

3.1.3 Conventional Methods for Analysing GWAS

GWAS require three key elements: (1) large volume of data which provides sufficient genetic information to answer certain research questions; (2) polymorphic alleles that can be genotyped efficiently with existing computing platforms and cover the whole genome adequately; (3) a statistical association testing method which can be used to conduct multiple comparisons and identify the genetic associations in an unbiased fashion.

Marginal tests have been used currently for testing the associations between SNPs and diseases. Due to a few limitations of marginal tests, various joint testing methods have been developed.

Variable or feature selection has become the important part of GWAS where datasets consist of hundreds of thousands of features. Feature selection improves the performance of predictors by providing more cost-effective predictors and thus provides a better knowledge of the underlying process and data [91].

Traditional ways of selecting variables include: forward, backward, stepwise (see Section 2.3.1). However, those procedures are unstable and computationally expensive even when the number of covariates is not large.

3.1.3.1 Marginal Testing

Marginal testing is also known as one at a time allele testing, univariate analysis and single SNP-based testing. The dependent variable y is a column of 1's and 0's representing cases and controls respectively and the independent variable X is a matrix consisting of 0's, 1's and 2's. Various methods have been tried to determine whether certain SNPs are associated with the disease. For binary traits in marginal testing, general methods of analyzing the data include logistic regression and contingency tables. Under the null hypothesis that there is no association between the phenotype and genotype, the χ^2 square test [92] or Fisher's exact test [93] can be employed to obtain a p-value and reject the hypothesis if

the p-value is below the significance level, which is normally set to be 0.05. For quantitative traits, generalized linear models or the analysis of variance (ANOVA) are normally used.

Significance level of 5% indicates that 5% of the time, the null hypothesis is rejected when in fact it is true and a false positive is reported. For the hundreds of thousands to millions of tests in GWAS, the cumulative likelihood of finding false positives is very large.

3.1.3.2 Multiple Testing

As the the number of hypotheses increases, the likelihood of finding a rare event rises, and therefore the chance of rejecting the null hypothesis when it is true increases. As a result, while a significance level α works fine for each individual test, it may not be appropriate for multiple tests. Therefore, the significance level α needs to be adjusted to account for multiple testing.

To correct for multiple testing and control the familywise error rate (FWER), one of the easiest and most common approaches, the Bonferroni adjustment, can be applied. Here, familywise error rate refers to the probability of having at least one false discovery in total when performing multiple hypotheses tests. The basic idea of the Bonferroni adjustment is to set the significance level of each individual test to be α divided by the total number of tests, say n, under the assumption that significance level for all tests is α. For example, if the total number of multiple hypotheses tests is 10, then the significance level set for each single test should be $\alpha = 0.05/10 = 0.005$.

Instead of controlling the FWER, another popular way of correcting for the multiple testing is to control the false discovery rate (FDR) [94], which is the expected proportion of incorrectly rejected null hypotheses.

Suppose there are m known hypotheses tests of which m_0 is the number of true hypotheses tests. Define the observable random variable R is the number of hypotheses that have been rejected and U, V, S, T are all unobservable random

	Declared non-significant	Declared significant	Total
H_0 true	U	V	m_0
H_1 true	T	S	$m - m_0$
Total	$m - R$	R	m

TABLE 3.1: Possible outcomes of testing multiple null hypothesis.

variables. The number of times that errors occurred when testing m null hypotheses is classified in Table 3.1.

Recall that the FWER is the probability of having at least one false discovery when performing multiple hypotheses tests and it is expressed as $P(V \geq 1)$. Hence $P(V \geq 1) \leq \alpha$ when significance level for each individual test is α/m.

The FDR Q_e is defined as the expectation of the random variable $Q = V/(V + S)$ which is the proportion of errors occurred by incorrectly rejecting the null hypotheses, and thus $E(Q) = E(V/R)$. It is obvious that when there is no rejection $R = 0$, there is no false rejection and hence $Q = 0$.

If all of the null hypotheses are true, then $S = 0$ and $V = R$, so $Q = 0$ when $V = 0$ and $Q = 1$ when $V > 0$. As a result, $E(Q) = 0 \cdot P(Q = 0) + 1 \cdot P(Q = 1) = P(Q = 1) = P(V > 0) = P(V \geq 1)$ which implies that controlling the FDR is the same as controlling the FWER.

If some of the null hypotheses are untrue, then $P(V \geq 1) \geq Q_e$ which implies that FDR controlling procedures is less stringent over false discovery than familywise error rate procedures, and thus has more power.

The procedure is as follows: firstly, order p-values of m total hypotheses tests; secondly, find the largest integer k such that the kth p-value $p_{(k)}$ is smaller than or equal to $k\alpha/m$; finally, reject any hypothesis test whose p-value is smaller than or equal to $p_{(k)}$.

Although marginal tests are easy to comprehend and conduct [95] and have detected a number of significant SNPs for various diseases [96], there are three major weaknesses:

1. Marginal tests might not be able to detect a large portion of genetic variation where diseases are very often polygenic, and may struggle to identify weaker association [97].

2. It is impossible to characterize genetic interactions of different genes.

3. They are not powerful in identifying gene-environment interactions [98].

3.1.4 Modern Statistical Methods for Analysing GWAS

For the case-control cohorts, simultaneous analysis of SNPs for GWAS has become more and more popular. Modern approaches have been developed for selection and estimation such as: ridge [8], bridge [99], Lasso [9], elastic net [36] (see Section 2.3).

The logistic regression can be used to fit all covariates simultaneously. However, since there are too many predictors in the model, there may exist over-fitting. An introduction of a proper prior or a penalty is needed to set some of β_j's to be 0. For example, a Lasso penalised logistic regression can be introduced [100], a Bayesian-inspired penalised maximum likelihood approach for variable selection in a logistic regression [97], a variational Bayes algorithm [101]. More recently, in high dimensional situations, a two-stage procedure for variable selection has been employed [89, 98, 102, 103].

3.1.4.1 The Bayesian Lasso for GWAS

Current widely used method in genome-wide association studies is univariate regression which analyses one SNP at a time by adjusting for multiple testing. This method has several disadvantages. It does not take interactions into consideration and it is not powerful in detecting weaker associations.

Li et al. [98] proposed a two-stage procedure for identifying important SNPs. The first stage is preconditioning via a supervised principal component analysis. Suppose the linear model is of the form: $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, then the design matrix $\mathbf{X}$

is reduced to $\mathbf{X}_r$, where $\mathbf{X}_r$ consists of the columns of $j' \in \{j : |\hat{b}_j| > \theta\}$ only. In practice, θ is selected using 5-fold cross-validation method.

At a second stage, a standard variable selection procedure such as the Lasso penalised regression can be conducted for the preconditioned response variable $\tilde{\mathbf{y}}$:

$$\tilde{y}_i = \mu + \mathbf{X}_i^T\boldsymbol{\alpha} + \mathbf{Z}_i^T\boldsymbol{\beta} + \boldsymbol{\xi}_i^T\mathbf{a} + \boldsymbol{\zeta}_i^T + \epsilon_i,\ i = 1, \cdots, n,$$

and the jth element of ξ_{ij} and ζ_{ij} are:

$$\xi_{ij} = \begin{cases} 1 & \text{if the genotpye of SNP j is AA} \\ 0 & \text{if the genotpye of SNP j is Aa} \\ -1 & \text{if the genotpye of SNP j is aa,} \end{cases} \quad \text{and}$$

$$\zeta_{ij} = \begin{cases} 1 & \text{if the genotype of SNP j is Aa} \\ 0 & \text{if the genotype of SNP j is AA or aa.} \end{cases}$$

The above equation can be rewritten as Lasso penalised least squares:

$$\frac{1}{2}||\tilde{\mathbf{y}} - \boldsymbol{\mu} - \mathbf{X}\boldsymbol{\alpha} - \mathbf{Z}\boldsymbol{\beta} - \boldsymbol{\xi}\mathbf{a} - \boldsymbol{\zeta}\mathbf{d}||^2 + \lambda\sum|a_j| + \lambda^*\sum|d_j|,$$

where λ and λ^* are tuning parameters that control the degrees of shrinkage in the estimate of additive and dominant effects. Then the Bayesian Lasso is implemented and solved using the Markov chain Monte Carlo (MCMC) algorithm.

Simulation results show that simultaneous analysis outperforms single SNP analysis. It also improves the probability of correctly identifying non-zero coefficients although it has slightly higher false positive rate. As the number of SNPs increases, the preconditioned Bayesian Lasso is still able to identify non-zero coefficients in almost every case where the traditional single SNP test fails.

The preconditioned Bayesian Lasso produces sparse solutions. Since the Bayesian Lasso is based on the Gibbs sampler, it provides point estimates as well as interval estimates of all parameters. It automatically chooses the tuning parameters and automatically accounts for the uncertainty in the estimation.

There are some drawbacks of this procedure. The computation time is relatively high. In the context of GWAS, marginally independent SNPs could be screened out by preconditioning while standard variable selection techniques, such as Lasso and Bayesian Lasso, are able to identify them. In addition, this approach may not work well if there is interactions between SNPs.

3.1.4.2 Sparse Group Lasso with Group Graph Structure for GWAS

Most recently, a two-level structures sparse model called sparse group Lasso with group-level graph structure (SGLGG) [104] is proposed. The model consists of two levels of predictors: the nucleotide level (SNPs) and the gene level. The SGLGG is able to select a number of important SNPs within each gene group and select a number of important gene groups on the entire sequence.

Suppose there are p predictors in total and there are K non-overlapped groups, then p_k denotes the number of low-level predictors in group k. Let $\mathbf{s}$ be the low-level predictors and $\mathbf{g}$ be the group-level predictors, then the low-level predictors can be expressed as:

$$\mathbf{s} = [s_{11}, \ldots, s_{1p_1}, \ldots, s_{k1}, \ldots, s_{kp_k}].$$

Let

$$\mathbf{G}_s = [g_1 s_{11}, \ldots, g_1 s_{1p_1}, \ldots, g_k s_{k1}, \ldots, g_k s_{kp_k}],$$

where g_j is the i-th element of $\mathbf{g}$. The SGLGG optimisation problem is proposed as:

$$\min_{\mathbf{g},\mathbf{s}} \ell(\mathbf{y}, \mathbf{G}_s) + \lambda_1 \|\mathbf{w}_\mathbf{g} \circ \mathbf{g}\|_1 + \lambda_2 \sum_{(i,j)\in E} (\tau(r_{ij})|g_i - \text{sign}(r_{ij})g_j|) + \lambda_3 \|\mathbf{s}\|_1, \quad (3.1)$$

where $\ell(\cdot)$ is a convex empirical loss function, $\circ$ is the Hadamard product operator, $\tau(r_{ij})$ is a monotonically increasing weight function which impose fusion between g_i and g_j. The SGLGG problem consists of three constraints, a group-level sparsity, a group-level graph structure via the fused Lasso and a low-level sparsity.

The minimisation problem in (3.1) can also be expressed as a simplified SGLGG model

$$\min_{\mathbf{g},\mathbf{s}} \ell(\mathbf{y}, \mathbf{G}_s) + \lambda_1 \|\mathbf{g}\|_1 + \lambda_2 \|\mathbf{T}\mathbf{g}\|_1 + \lambda_3 \|\mathbf{s}\|_1, \tag{3.2}$$

where $\mathbf{T}$ is the sparse matrix constructed from the edge set E, $t_{ij} = t_{ji} = r_{ij}$ if there is an edge between g_i and g_j and r_{ij} is the weight of the edge between node g_i and g_j. However, this model is hard to solve directly, and therefore the original problem can be reformulated using constrained optimization methods. This optimization problem can be solved using the alternating direction method of multipliers [42] as discussed in Section 2.3.4.4.

The ADMM for SGLGG problem can be written as

$$\begin{aligned} &\min_{\mathbf{g},\mathbf{s},\mathbf{p},\mathbf{q},\mathbf{r}} \frac{1}{2}\|\mathbf{y} - \mathbf{A}\mathbf{G}_\mathbf{s}\|^2 + \lambda_1\|\mathbf{p}\|_1 + \lambda_2\|\mathbf{q}\|_1 + \lambda_3\|\mathbf{r}\|_1 \\ &\text{subject to } \mathbf{g} - \mathbf{p} = 0, \mathbf{T}\mathbf{g} - \mathbf{q} = 0, \mathbf{s} - \mathbf{r} = 0 \end{aligned}$$

The augmented Lagrangian is then expressed as

$$\begin{aligned} L_\rho(\mathbf{g}, \mathbf{s}, \mathbf{p}, \mathbf{q}, \mathbf{r}, \mathbf{u}, \mathbf{v}, \mathbf{w}) = &\frac{1}{2}\|\mathbf{y} - \mathbf{A}\mathbf{G}_\mathbf{s}\|^2 + \lambda_1\|\mathbf{p}\|_1 + \lambda_2\|\mathbf{q}\|_1 + \lambda_3\|\mathbf{r}\|_1 + \\ &\mathbf{u}^\top(\mathbf{g} - \mathbf{p}) + \mathbf{v}^\top(\mathbf{T}\mathbf{g} - \mathbf{q}) + \mathbf{w}^\top(\mathbf{s} - \mathbf{r}) + \\ &\frac{\rho}{2}\|\mathbf{g} - \mathbf{p}\|^2 + \frac{\rho}{2}\|\mathbf{T}\mathbf{g} - \mathbf{q}\|^2 + \frac{\rho}{2}\|\mathbf{s} - \mathbf{r}\|^2, \end{aligned}$$

where $\mathbf{u}$, $\mathbf{v}$, $\mathbf{w}$ are Lagrangian multipliers.

For step $k = 1, 2, \ldots$, the variables are updated as:

$$
\begin{aligned}
\mathbf{g}^{(k+1)} &:= \arg\min_{\mathbf{g}} L_\rho(\mathbf{g}, \mathbf{s}^{(k)}, \mathbf{p}^{(k)}, \mathbf{q}^{(k)}, \mathbf{r}^{(k)}, \mathbf{u}^{(k)}, \mathbf{v}^{(k)}, \mathbf{w}^{(k)}) \\
\mathbf{s}^{(k+1)} &:= \arg\min_{\mathbf{s}} L_\rho(\mathbf{g}^{(k+1)}, \mathbf{s}, \mathbf{p}^{(k)}, \mathbf{q}^{(k)}, \mathbf{r}^{(k)}, \mathbf{u}^{(k)}, \mathbf{v}^{(k)}, \mathbf{w}^{(k)}) \\
\mathbf{p}^{(k+1)} &:= \arg\min_{\mathbf{p}} L_\rho(\mathbf{g}^{(k+1)}, \mathbf{s}^{(k+1)}, \mathbf{p}, \mathbf{q}^{(k)}, \mathbf{r}^{(k)}, \mathbf{u}^{(k)}, \mathbf{v}^{(k)}, \mathbf{w}^{(k)}) \\
\mathbf{q}^{(k+1)} &:= \arg\min_{\mathbf{q}} L_\rho(\mathbf{g}^{(k+1)}, \mathbf{s}^{(k+1)}, \mathbf{p}^{(k+1)}, \mathbf{q}, \mathbf{r}^{(k)}, \mathbf{u}^{(k)}, \mathbf{v}^{(k)}, \mathbf{w}^{(k)}) \\
\mathbf{r}^{(k+1)} &:= \arg\min_{\mathbf{r}} L_\rho(\mathbf{g}^{(k+1)}, \mathbf{s}^{(k+1)}, \mathbf{p}^{(k+1)}, \mathbf{q}^{(k+1)}, \mathbf{r}, \mathbf{u}^{(k)}, \mathbf{v}^{(k)}, \mathbf{w}^{(k)}) \\
\mathbf{u}^{(k+1)} &:= \mathbf{u}^{(k)} + \rho(\mathbf{g}^{(k+1)} - \mathbf{p}^{(k+1)}) \\
\mathbf{v}^{(k+1)} &:= \mathbf{v}^{(k)} + \rho(\mathbf{T}^{(k+1)}\mathbf{g}^{(k+1)} - \mathbf{q}^{(k+1)}) \\
\mathbf{w}^{(k+1)} &:= \mathbf{w}^{(k)} + \rho(\mathbf{s}^{(k+1)} - \mathbf{r}^{(k+1)}).
\end{aligned}
\tag{3.3}
$$

The proposed method was applied to the Alzheimer's disease (AD) genome sequence data and neuroimage data. The experiment results show that SGLGG is very competitive in terms of the predictive performance of the AD-related phenotypes, compared with other state-of-the-art sparse learning models. In addition, the SGLGG is very efficient in identifying risk SNPs with respect to AD imaging phenotypes. With the help of gene-level biological priors, SGLGG has good prospects for revealing SNP–SNP interactions among different genes.

3.1.5 The Significance of Statistical Analysis to GWAS

Genome-wide association studies have had a significant influence in human genetics. Although there are a number of risk factors that have been found for some common disease, more are expected to be identified. With hundreds of thousands of SNPs and only a few cases and controls compared to the SNPs, it is still a challenge for people to develop more advanced techniques to handle massive amount of data and identify the relationship between genotypes and phenotypes.

Therefore, the future of human genetics heavily relies on the incorporation of statistical methods to complex genetics data. The potential significance of statistical

analysis to GWAS this is to improve healthcare such as enabling personalised medicine and development of new strategies to fight disease, improving prognosis, identifying novel disease pathways and therapeutic targets, encouraging collaborative endeavor and advancing knowledge of genetic architecture.

3.2 Breast Cancer

3.2.1 Background

Cancer is one of the biggest threats for human health. Among different types of cancers, breast cancer ranks as the fifth cause of death from cancer and over one million new cases of breast cancer diagnosed every year. In the less developed countries, it account 14.3% of cancer-related deaths. The impact is more significant in developing countries, it is the second cause of death from cancer and accounts for 15.4% of cancer-related deaths in women [105].

Diagnose at an advanced stage causes difficulty of tumor removal. The early stage detection would be helpful to give patients more effective treatment and increase the chance of survival. Therefor, an assessment model to detect people with high risk of breast cancer at the early stages would be greatly appreciated by patients, clinicians, and researchers.

Breast cancer can be separated into different types based on genetic variations. Different treatments will be used for each sub-type which leads to distinct clinical outcomes. There are three clinical types of breast cancer based on tumor histological biomarkers, estrogen receptor (ER) and progesterone receptor (PR) positive, human epidermal growth factor receptor 2 positive (HER2+), and triple-negative (ER-/PR-/HER2-) breast cancer [106]. About 80% of all breast cancers are ER-positive and about 65% of the ER-positive cases are also PR-positive [107, 108], about 20% are HER2-positive [109] and nearly 15% are triple-negative [110].

Around 10% of women with breast cancer have a family history of the disease which is an established risk factor for breast cancer [111, 112]. Early studies used linkage analysis and positional cloning in families with multiple affected individuals to identify genetic factors associated with breast cancer predisposition [113, 114]. Previous studies discovered genes such as *BRCA1* [113] and *BRCA2* [114]. *BRCA1* and *BRCA2* can be found in about 25% of the breast cancer families [115, 116]. The other high-risk breast-cancer genes include *PTEN* and *TP43* which are relatively rare [117–119].

Risk prediction models have been widely used to identify individuals with high risk of breast cancer. For example, the Gail model [120] has been used to estimate the absolute risk of invasive breast cancer for individuals. This model has been widely applied in areas such as clinical counselling and breast cancer prevention studies. Risk factors such as family history, age and location have been used to predict breast cancer. Table 3.2 shows some traditional risk factors of breast cancer [1].

However, the major contributor of identifying risk predictors over past decade was genome-wide associations studies (GWAS). The GWAS is based on genome-wide genotyping for thousands to millions of single-nucleotide polymorphisms in a large

Factor		Relative Risk	High Risk Group
Age		> 10	Elderly
Geographical location		5	Developed country
Age at menarche		3	Menarche before age 11
Age at menopause		2	Menopause after age 54
Age at first full pregnancy		3	First child in early 40s
Family history		≥ 2	Breast cancer in first degree relative when young
Previous benign disease		4-5	Atypical hyperplasia
Cancer in other breast		> 4	
Socioeconomic group		2	Groups I and II
Diet		1.5	High intake of saturated fat
Body weight:	Premenopausal	0.7	Body mass index > 35
	Postmenopausal	2	Body mass index > 35
Alcohol consumption		1.3	Excessive intake
Exposure to ionising radiation		3	Abnormal exposure in young females after age 10
Taking exogenous hormones:	Oral contraceptives	1.24	Current use
	Hormone replacement therapy	1.35	Use for ≥ 10 years
	Diethylstillbestrol	2	Use during pregnancy

TABLE 3.2: Reproduction of Non-genetic Risk factors for breast cancer [1].

scale of individuals and contrast between groups of individuals with and without a certain phenotype.

3.2.2 GWAS for Breast Cancer

Over the past decade, GWAS has been widely applied to identify a large number of novel asscoications. Various susceptibility SNPs have been identified in large-scale and different populations by researchers and global consortium groups such as the Asia Breast Cancer Consortium (ABCC) and Breast Cancer Association Consortium (BCAC)

In 2007, the first GWAS of breast cancer identified five novel independent loci that have strong association with breast cancer [121]. The genes around four of the novel loci are plausible causative genes (*FGFR2*, *TNRC9* at 16q locus, *MAP3K1* at 5q locus and *LSP1* at 11p locus). The fifth locus is an interval of 110kb lacking known genes. The most strong SNP has been identified in the intron 2 of the *FGFR2* gene. Soon after, four SNPs in intron 2 of *FGFR2* were identified and confirmed the association of *FGFR2* gene with breast cancer [122]. Later on, Stacey et al. found four SNPs consistently associated with breast cancer with two of them at locus 2q and 16q [123], and another two at 5p locus [124]. An additional locus at 6q which include genes *ECHDC1* and *RNF146* was found [125]. Moreover, strong evidence showed that susceptibility loci at 3p and 17q with potential causative genes include *SLC4A7* and *NEK10* on 3p and *COX11* on 17q [126]. Thomas et al. conducted a three-stage GWAS and found two SNPs at two novel loci respectively, 1p and 14q locus [127]. The first SNP resides in a large linkage disequilibrium block neighboring genes *NOTCH2* and *FCGR1B*, and the second SNP localised to gene *RAD51L1* which is a gene in the homologous recombination DNA repair pathway. Zhang et al. conducted a further GWAS on Chinese women and found a SNP at a susceptibility locus, 6q, located upstream of the gene encoding estrogen receptor alpha (*ESR1*) which showed strong and consistent association with breast cancer [128].

In 2010, five new susceptibility loci on chromosome 9, 10 and 11 were found, and three SNPs at locus $6q$, $8q$ and $11p$ showed most significant association with risk than those reported previously [129]. In the same year, Antoniou et al. [130] found five SNPs at locus $19p$ that were associated with breast cancer risk.

In 2011, new loci associated with breast cancer has been discovered. Haiman et al. [131] combined GWAS data from African and European women to research on the common risk alleles for ER-negative breast cancer. Their results identified a risk variant at the *TERT-CLPT1L* locus on chromosome $5p15$ which was also significantly associated with triple-negative breast cancer, particularly among women younger than 50 years old. Cai et al. [132] conducted a four-stage GWAS including 17153 cases and 16943 controls among Chinese, Korean and Japanese women and their results showed strong implication that one SNP at $10q21.2$ is a genetic risk variant for breast cancer.Fletcher et al. [133] conducted a GWAS and identified a locus risk for breast cancer at $9q31.2$ and two variants mapping to $6q25.1$ estrogen receptor 1, which were assiciated with breast cancer among women with northern European ancestry.

In 2012, Long et al. [134] conducted a four-stage GWAS with data of Chinese, Korean and Japanese women to discover genetic susceptibility loci for breast cancer. Their results identified a SNP in chromosome $6q25.1$ as a consistent association with breast cancer risk across all four stages, and also another two possible susceptibility loci located in the *ESR1* gene and 11q24.3.

Kim et al. [135] conducted a three-stage GWAS on Korean women and identified seven breast cancer susceptibility loci: 5q11.2/*MAP3K1* (rs889312 and rs16886165), 5p15.2/*ROPN1L* (rs1092913), 5q12/*MRPS30* (rs7716600), 6q25.1/*ESR1* (rs2046210 and rs3734802), 8q24.21 (rs1562430), 10q26.13/*FGFR2* (rs10736303), and 16q12.1/*TOX3* (rs4784227 and rs3803662). They were significantly associated with breast cancer risk in Korean women. In addition, they also found SNP rs13393577 at chromosome 2q34 as a consistent association with breast cancer risk. Siddiq et al. [136] conducted a analysis of ER-negative disease in Japanese, Latino and European ancestry. They identified two loci for breast cancer at 20q11 and

6q14. SNP rs2284378 at 20q11 which were associated with ER-negative breast cancer.

3.3 Summary

In this chapter, we first introduced basic concepts of genetics. Next, we introduced genome-wide association studies, conventional approaches as well as recent developments for analysing GWAS. Then, we discussed background of breast cancer such as different types of breast cancer, risk factors, its impact, etc. Then we also reviewed GWAS for breast cancer, the methodology and some current findings based on GWAS.

Chapter 4

Bayesian Grouped Horseshoe Models with Application to Additive Models

4.1 Introduction

Statistical variable selection, also known as feature selection, has become an indispensable tool in many research areas involving machine learning and data mining. The object is to select the best subset of predictors for fitting or predicting the response variable from a potentially large collection of candidate predictors. It is particularly important in high-dimensional problems, where there are potentially hundreds, or even thousands, of predictors and only a few are associated with the outcome.

Consider the linear regression model:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{4.1}$$

where $\mathbf{y}$ is an n by 1 observation vector of the response variable, $\mathbf{X}$ is an n by p observation or design matrix of the regressors or predictors, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ is a

p by 1 vector of regression coefficients to be estimated, and $\boldsymbol{\epsilon}$ is an n by 1 vector of i.i.d. $\mathcal{N}(0, \sigma^2)$ random errors with σ^2 unknown. Here, the vector $\boldsymbol{\beta}$ is assumed to be sparse in the sense that most of its components are equal to zero. Therefore, dimensionality reduction is necessary, especially for large-p problems.

Recent approaches for variable selection in ultra-high dimensions are based on penalised likelihood methods, which select a model by minimising a loss function that is usually proportional to the negative log likelihood plus a penalty term:

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta} \in \mathbb{R}^p}{\arg\min} \left\{ (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + \kappa \cdot q(\boldsymbol{\beta}) \right\}, \tag{4.2}$$

where $\kappa > 0$ is the regularisation parameter and $q(\cdot)$ is a penalty function. Many penalised regression approaches have been proposed in the literature, such as the non-negative garrote [7], ridge regression [8], the adaptive lasso [11], the elastic net [36], and the smoothly clipped absolute deviation [45]. One of the most widely used penalised approaches is the lasso [59], which shrinks coefficients while setting other coefficients to zero, effectively producing a model that is simpler to interpret. However, penalised regression methods mentioned above, such as the lasso discussed in Section 2.4.2.2, are designed for selecting individual explanatory variables.

4.1.1 Grouped Variables

Group structures naturally exist in predictor variables, and in many situations, groups of variables are desired to be selected rather than individual variables. For example, a multi-level categorical predictor in a regression model can be represented by a group of dummy variables, a continuous predictor in an additive model can be expressed as a composition of basis functions, and prior knowledge such as genes in the same biological pathway can be used to form a natural group.

To find sparse solutions at the group level, several selection methods have been proposed. These include the group lasso [37], the group lasso for logistic regression

[137], group selection with general composite absolute penalty [138], the group bridge method [139], and the bi-level selection in generalised linear models [140].

4.1.2 Bayesian Regression

An important class of alternative variable selection methods are the Bayesian penalised regression approaches (see Section 2.4). These are motivated by the fact that a good solution for β in (4.1) can be interpreted as the posterior mode of β in the Bayesian model when β follows a certain prior distribution. For example, the assumption of a Laplace prior distribution for β leads to the Bayesian interpretation of the lasso [59].

A natural way of estimating β in the Bayesian approaches is generating sparsity through a *spike and slab* prior for each element of β, with the spike concentrating near zero and the slab spreading away from zero:

$$\beta_j|\gamma_j \sim (1-\gamma_j)\mathcal{N}(0,\tau_j^2) + \gamma_j\mathcal{N}(0,c_j\tau_j^2),\ \ j=1,\ldots,p, \tag{4.3}$$

where γ_j are binary variables, τ_j^2 is small and $c_j > 0$ is large. A fully Bayesian approach with a spike and slab mixture for each β requires exploration of a model space of size 2^p, which becomes difficult when p is large.

Another way of estimating β in the Bayesian approaches is via shrinkage priors, such as the Bayesian lasso. A competitive Bayesian alternative to the lasso is the Bayesian model resulting from the horseshoe prior (see Section 2.4.2.3), which is a one-component prior [65]. The horseshoe arises from the same class of multivariate scale mixtures of normals as the lasso, but it is generally superior to the double-exponential prior at handling sparsity [62]. Furthermore, the horseshoe prior is known for its robustness at handling large outlying signals and the estimator of the horseshoe model does not face the computational issues of the point mass mixture models. To date, grouped variable selection methods based on the Bayesian horseshoe (BHS) model have not been analysed in the literature.

In the following section, we will extend the BHS model by introducing shrinkage parameters at the group level as well as within each group to handle grouped variables. The proposed models have the advantages of handling sparsity as well as strong signals when grouped structures are presented in the data. We will then apply the proposed grouped models to the important class of additive models where group structures naturally exist. Both simulated and real data experiments demonstrate the promising performance of the proposed methods.

4.2 Bayesian Group Horseshoe Models

In this section, the BHS model is briefly introduced and two extensions for grouped variables are proposed.

4.2.1 Bayesian Horseshoe Model

The horseshoe prior assumes that β_j are conditionally independent and each has a density function that can be represented as a scale mixture of normals. The horseshoe prior leaves strong signals unshrunk and penalises noise variables severely. Therefore, the horseshoe prior has the ability to adapt to different sparsity patterns, while simultaneously avoiding over-shrinkage of large coefficients.

Without loss of generality, $\mathbf{y}$ and $\mathbf{X}$ are assumed to be standardised for all models. The response $\mathbf{y}$ is centered, and the covariates $\mathbf{X}$ are column-standardised to have mean zero and unit length. The BHS can be expressed as

$$\begin{aligned}
\mathbf{y}|\mathbf{X}, \boldsymbol{\beta}, \sigma^2 &\sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n) \\
\boldsymbol{\beta}|\sigma^2, \tau^2, \lambda_1, \ldots, \lambda_p &\sim \mathcal{N}(\mathbf{0}, \sigma^2\tau^2\mathbf{D}_{\boldsymbol{\lambda}}), \text{ where } \mathbf{D}_{\boldsymbol{\lambda}} = \text{diag}(\lambda_1^2, \ldots, \lambda_p^2) \\
\lambda_j &\sim C^+(0,1), \ j = 1, \ldots, p \\
\tau &\sim C^+(0,1) \\
\sigma^2 &\sim \frac{1}{\sigma^2} d\sigma^2,
\end{aligned} \tag{4.4}$$

where $C^+(0,1)$ is a standard half-Cauchy distribution with the probability density function:

$$f(x) = \frac{2}{\pi(1+x^2)}, \ x > 0. \tag{4.5}$$

The scale parameters λ_j are local shrinkage parameters and τ is the global shrinkage parameter. A simple sampler proposed for the BHS hierarchy [67] enables straightforward sampling of the full conditional posterior distributions.

4.2.2 Bayesian Grouped Horseshoe Model

The original BHS model does not allow for grouped structures. Therefore, the Bayesian grouped horseshoe (BGHS) model is proposed. Suppose there are p predictors, $G \in [1,p]$ groups of predictors in the data and the gth group has size s_g, where $g = 1,\ldots,G$ (i.e. there are s_g variables in group g). The horseshoe hierarchical representation of the full model for grouped variables can be constructed as:

$$\begin{aligned}
\mathbf{y}|\mathbf{X},\boldsymbol{\beta},\sigma^2 &\sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta},\sigma^2\mathbf{I}_n) \\
\boldsymbol{\beta}|\sigma^2,\tau^2,\delta_1,\ldots,\delta_G &\sim \mathcal{N}(\mathbf{0},\sigma^2\tau^2\mathbf{D}_{\boldsymbol{\delta}}), \text{ where } \mathbf{D}_{\boldsymbol{\delta}} = \text{diag}(\delta_1^2\mathbf{I}_{s_1},\ldots,\delta_G^2\mathbf{I}_{s_G}) \\
\delta_g &\sim C^+(0,1), \ g = 1,\ldots,G \\
\tau &\sim C^+(0,1) \\
\sigma^2 &\sim \frac{1}{\sigma^2}d\sigma^2,
\end{aligned} \tag{4.6}$$

where δ_g are the shrinkage parameters at group level. Instead of having shrinkage parameters for all predictors, we have shrinkage parameters δ_g for each group. Hence, the model either shrinks all variables in the same group towards zero or leaves them untouched.

4.2.3 Hierarchical Bayesian Grouped Horseshoe Model

By combining the original BHS model with the BGHS model, we develop a hierarchical Bayesian grouped horseshoe (HBGHS) model that does selection and

shrinkage at the group level as well as within groups.

The total number of groups G is assumed to be greater than one because $G = 1$ implies that all predictors are in the same group and results in the regular BHS model. The full HBGHS model is:

$$\begin{aligned}
\mathbf{y}|\mathbf{X}, \boldsymbol{\beta}, \sigma^2 &\sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n) \\
\boldsymbol{\beta}|\sigma^2, \tau^2, \delta_1, \ldots, \delta_G, \lambda_1, \ldots, \lambda_p &\sim \mathcal{N}(\mathbf{0}, \sigma^2\tau^2\mathbf{D}_{\boldsymbol{\delta}}\mathbf{D}_{\boldsymbol{\lambda}}) \\
\text{where } \mathbf{D}_{\boldsymbol{\delta}} = \text{diag}(\delta_1^2 I_{s_1}, \ldots, \delta_G^2 I_{s_G}) \quad & \mathbf{D}_{\boldsymbol{\lambda}} = \text{diag}(\lambda_1^2, \ldots, \lambda_p^2) \\
\delta_g &\sim C^+(0,1),\ g = 1, \ldots, G \\
\lambda_j &\sim C^+(0,1),\ j = 1, \ldots, p \\
\tau &\sim C^+(0,1) \\
\sigma^2 &\sim \frac{1}{\sigma^2} d\sigma^2,
\end{aligned} \tag{4.7}$$

where $\lambda_1, \ldots, \lambda_p$ are the shrinkage parameters for each predictor variable and $\delta_1, \ldots, \delta_G$ are the shrinkage parameters for the group variables. Therefore, the model has a global shrinkage parameter τ, and local shrinkage parameters δ and λ, which control shrinkage at the group level and within group levels, respectively.

The conditional distribution of $\boldsymbol{\beta}$ is

$$\boldsymbol{\beta}|\sigma^2, \tau^2, \delta_1^2, \ldots, \delta_G^2, \lambda_1^2, \ldots, \lambda_p^2 \sim \mathcal{N}\left(\mathbf{A}^{-1}\mathbf{X}^T\mathbf{y}, \sigma^2\mathbf{A}^{-1}\right),$$

where

$$\mathbf{A} = \mathbf{X}^T\mathbf{X} + (\tau^2\mathbf{D}_{\boldsymbol{\delta}}\mathbf{D}_{\boldsymbol{\lambda}})^{-1}.$$

The full conditional distribution of σ^2 is

$$\sigma^2|\boldsymbol{\beta}, \tau^2, \delta_1^2, \ldots, \delta_G^2, \lambda_1^2, \ldots, \lambda_p^2 \sim \mathcal{IG}\left(\frac{n-1+p}{2}, \frac{(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^T(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) + \boldsymbol{\beta}^T(\tau^2\mathbf{D}_{\boldsymbol{\delta}}\mathbf{D}_{\boldsymbol{\lambda}})^{-1}\boldsymbol{\beta}}{2}\right).$$

The full conditional distribution of group shrinkage parameter δ_g^2 is

$$\delta_g^2|\boldsymbol{\beta}, \sigma^2, \tau^2, t_g, \lambda_1^2, \ldots, \lambda_p^2 \sim \mathcal{IG}\left(\frac{s_g+1}{2}, \frac{\boldsymbol{\beta}_g^T(\mathbf{D}_{\boldsymbol{\lambda}_g})^{-1}\boldsymbol{\beta}_g}{2\sigma^2\tau^2} + \frac{1}{t_g}\right),\ t_g|\delta_g^2 \sim \mathcal{IG}\left(1, \frac{1}{\delta_g^2}+1\right).$$

The full conditional distribution of the local shrinkage parameter λ_j^2 is

$$\lambda_j^2|\boldsymbol{\beta}, \sigma^2, \tau^2, \delta_1^2, \ldots, \delta_G^2, c_j \sim \mathcal{IG}\left(1, \frac{\beta_j^2}{2\sigma^2\tau^2\delta_{gj}^2} + \frac{1}{c_j}\right),\ c_j|\lambda_j^2 \sim \mathcal{IG}\left(1, \frac{1}{\lambda_j^2}+1\right).$$

The full conditional distribution of the global shrinkage parameter τ^2 is

$$\tau^2|\boldsymbol{\beta}, \sigma^2, \tau^2, \delta_1^2, \ldots, \delta_G^2, \lambda_1^2, \ldots, \lambda_p^2, v \sim \mathcal{IG}\left(\frac{p+1}{2}, \frac{\boldsymbol{\beta}^T(\mathbf{D}_{\boldsymbol{\delta}}\mathbf{D}_{\boldsymbol{\lambda}})^{-1}\boldsymbol{\beta}}{2\sigma^2} + \frac{1}{v}\right).$$

The full conditional distribution of hyper-parameter v is

$$v|\tau^2 \sim \mathcal{IG}\left(1, \frac{1}{\tau^2}+1\right).$$

4.3 Additive Models

Nonparametric regression methods, such as the additive model, are widely used in statistics. In an additive model, each predictor can be expressed as a set of basis functions that form a group structure. Given a data set $\{y_i, x_{i1}, \ldots, x_{ip}\}_{i=1}^n$, the additive model has the form:

$$y = \mu_0 + \sum_{j=1}^{p} f_j(X_j) + \epsilon, \tag{4.8}$$

where μ_0 is an intercept term and $f_j(\cdot)$ are unknown smooth functions. In the ideal situation, estimation of the selected smooth function is expected to be as close to the corresponding true underlying functions or target functions as possible.

There are various classes of basis functions that can be used to approximate the target functions. The basis functions include polynomials, spline functions, etc.

4.3.1 Polynomial Expansions

Let $g_j(x), j = 1, \ldots, p$, be a set of basis functions. Each smooth function component in the additive model can be represented as:

$$f(x) = a_0 + a_1 g_1(x) + a_2 g_2(x) + \cdots + a_p g_p(x). \tag{4.9}$$

The regular polynomial approximations can be expressed as:

$$f(x) = a_0 + a_1 x + a_2 x^2 + \cdots + a_p x^p. \tag{4.10}$$

One limitation of the regular polynomial basis functions (i.e. $g_j(x) = x^{j-1}$) is that there is potentially a high degree of correlation between the data generated by the basis functions. As a result, orthogonal polynomials are widely used. The formal definition of the orthogonal polynomial is that the basis functions that satisfy:

$$\int_a^b g_i(x) g_j(x) w(x) dx = 0, \quad \forall i, j, i \neq j, \tag{4.11}$$

where $w(\cdot)$ is the weighting function.

There are many orthogonal polynomials such as Chebyshev polynomials, Gegenbauer polynomial, Hermite polynomials, Jacobi polynomials, Legendre polynomials, etc. In this section, one of the orthogonal polynomial groups, the Legendre polynomials, are used for all polynomial expansions. The Legendre polynomials are defined on the interval $[-1, 1]$. Each Legendre polynomial $g_j(x)$ is a pth-degree polynomial and can be expressed as:

$$g_j(x) = \frac{1}{2^j j!} \frac{d^j}{dx^j} [(x^2 - 1)^j], \ p = 0, 1, \ldots. \tag{4.12}$$

The additive models allow for nonlinear effects and grouped structures, and therefore, form perfect test functions for our proposed methods.

4.3.2 Spline Expansions

Another popular function approximation method is called splines where the domain of the data is split into different lower order regions and each region is approximated using a different polynomial. Splines can be written as a combination of basis functions which are zero over most of the domain. The interpolating function can be expressed as:

$$S(x) = \begin{cases} S_0(x) & x \in [x_0, x_1); \\ S_1(x) & x \in [x_1, x_2); \\ \dots & \\ S_{n-1}(x) & x \in [x_{n-1}, x_n], \end{cases}$$

where x_i's are called knots through which spline functions pass.

The simplest spline is the piecewise linear interpolation where $S_i(x)$ are linear polynomials. However, piecewise linear splines are in general not differentiable at the knots. Another widely used spline expansions is the cubic splines which is constructed using piecewise third-order polynomials and is of the form:

$$S_i(x) = a_i + b_i x + c_i x^2 + d_i x^3, \text{ for } x \in [x_{i-1}, x_i],$$

where the second derivative of each polynomial is commonly set to zero at knots.

4.3.3 Fourier Expansions

A Fourier series is an expansion consisting of a series of sine and cosine functions, and it makes use of the orthogonality relationships of the sine and cosine functions [141, 142]. The Fourier expansion is particularly useful in obtaining solutions for an arbitrary periodic functions. It works by decomposing any periodic function into a set of simple functions where those simple functions can be solved separately and recombined to obtain approximate solutions to the original problem.

The Fourier expansion is of the form:

$$F_n(x) = a_0 + (a_1 \cos(x) + b_1 \sin(x)) + \cdots + (a_n \cos(nx) + b_n \sin(nx)),$$

where a_0 is the coefficient of the constant term, a_i and b_i are coefficients of $F_n(x)$, and the Fourier series is defined on the interval $[-\pi, \pi]$. Here the coefficients can be derived as:

$$\begin{aligned} a_0 &= \frac{1}{2\pi} \int_{-\pi}^{\pi} F_n(x) dx \\ a_j &= \frac{1}{\pi} \int_{-\pi}^{\pi} F_n(x) \cos(jx) dx, \quad j \in [1, n] \\ b_j &= \frac{1}{\pi} \int_{-\pi}^{\pi} F_n(x) \sin(jx) dx, \quad j \in [1, n]. \end{aligned}$$

4.3.4 Haar Wavelets

Another widely known basis functions that can be used to represent other functions are wavelets. Wavelets can be used to track information of both time and frequency. One of the simplest wavelets is the Haar wavelet [143]. It is known that any continuous function can be approximated uniformly by Haar functions.

There are two primary functions in wavelet analysis: the scaling function $\phi(x)$ and the wavelet $\psi(x)$, which are also known as the father wavelet and the mother wavelet respectively. The Haar scaling function (father wavelet) $\phi(x)$ is defined as:

$$\phi(x) = \begin{cases} 1, & \text{if } 0 \leq x < 1 \\ 0, & \text{otherwise.} \end{cases}$$

The Haar wavelet function (mother wavelet) $\psi(x)$ is defined as:

$$\psi(x) = \phi(2x) - \phi(2x - 1),$$

or in an equivalent notation:

$$\psi(x) = \begin{cases} 1 & \text{if } 0 \leq x < \frac{1}{2} \\ -1 & \text{if } \frac{1}{2} \leq x < 1 \\ 0 & \text{otherwise.} \end{cases}$$

The Haar basis consists of a set of functions:

$$\psi_{jk}(x) = 2^{j/2}\phi(2^j x - k),$$

where j is a non-negative integer and k is an interger. As a result, any function $f(x)$ represented using a Haar expansion can be expressed as:

$$f(x) = a_0 + \sum_{j=0}^{\infty} \sum_{k=0}^{2^j - 1} a_{jk}\psi_{jk}(x),$$

for $0 \leq k \leq 2^j - 1$.

The Haar wavelet processes several properties. Firstly, any continuous real function with compact support can be approximated uniformly by linear combinations of $\phi(x), \phi(2x), \ldots, \phi(2^n x), \ldots$ and their shifted functions. This extends to those function spaces where any function can be approximated by continuous functions. Secondly, any continuous real function on the interval $[0, 1]$ can be approximated uniformly on $[0, 1]$ by linear combinations of the constant function $1, \psi(x), \psi(2x), \psi(4x), \ldots, \psi(2^n x), \ldots$ and their shifted functions. Thirdly, the set of functions $\{2^{j/2}\phi(2^j x - k), k \in \mathbb{Z}\}$ is an orthonormal basis.

Figure 4.1 shows an example of basis functions for Haar wavelet expansions of degree 2. The basis functions consist of a series of discrete step functions. Therefore, the Haar wavelet expansions are expected to perform well when the true functions is a step function.

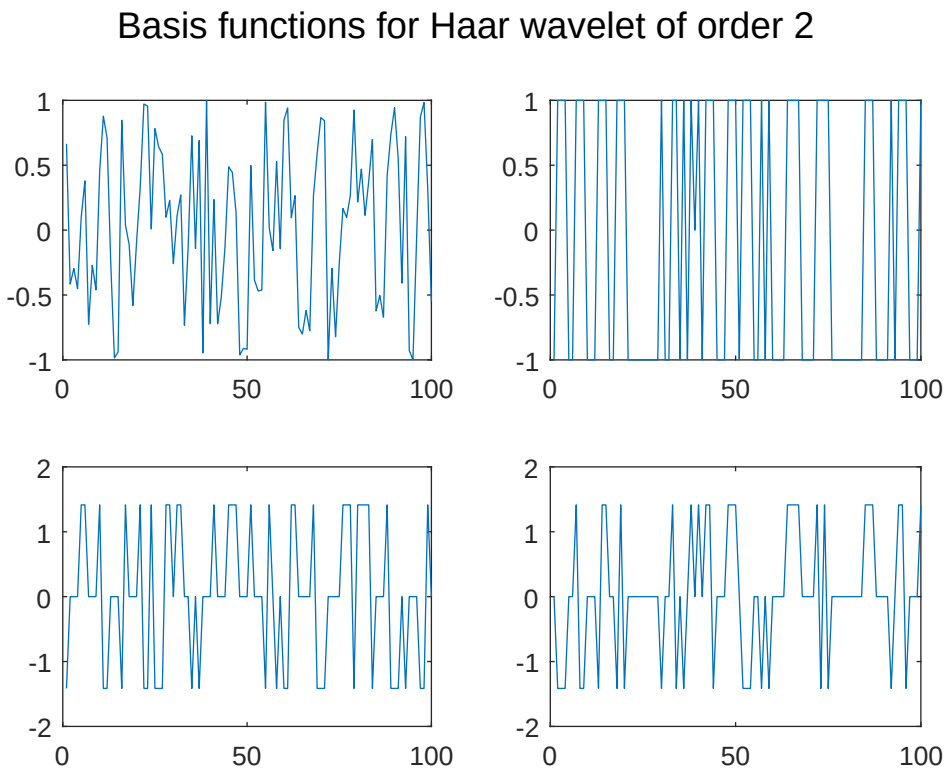

FIGURE 4.1: An example of various basis functions for Haar wavelet of degree $K = 2$.

4.4 Simulation Studies

In this section, we compared the prediction performance of four methods: (i) the regular BHS, (ii) the HBHS, (iii) lasso with regularisation parameter selected by the Bayesian information criterion (lasso-BIC) and (iv) the regular Bayesian horseshoe without expansion of the predictors (BHS-NE). The performances of the four methods were compared on three test functions (see Section 4.4.2).

To compare the performances of the four methods, we generated $\mathbf{X}$ with a large sample size n from a uniform distribution $U(-1, 1)$ and computed the mean squared prediction error (MSPE) as the comparison metric:

$$\frac{1}{n}\sum_{i=1}^{n}[\mathrm{E}(y_i|x_i) - \hat{y}_i]^2, \tag{4.13}$$

where y_i are the responses of the test functions and $\hat{y}_i$ are the fitted values based on the estimates from the four methods.

4.4.1 Simulation Procedures

For simulated data, we generated $t = 100$ data sets including $p = 10$ candidate predictors $\mathbf{X}_{n\times p}$ from the uniform distribution $U(-1, 1)$. For each realisation, we calculated $\mathbf{y}$ based on the true test functions. Then, we expanded $\mathbf{X}$ using Legendre polynomials with degree K. Three methods, BHS, HBGHS, and lasso-BIC, were applied to t samples of expanded data $\tilde{\mathbf{X}}$. The benchmark method, BHS-NE, which is the regular BHS applied to non-expanded covariates $\mathbf{X}$, was also included in the test for comparison. We obtained t posterior samples $\boldsymbol{\beta}$ for each method and computed the posterior means as the point estimates of each method.

4.4.2 Test Functions

Simulations were performed on data generated from the three true test functions shown below. The first true function is linear, the second is non-linear and the third consists of Legendre polynomials.

1. Test Function 1 (simple linear function):

$$y = X_1 + X_2 - X_3 - X_4 \tag{4.14}$$

2. Test Function 2 (nonlinear function):

$$y = \cos(8X_1) + X_2^2 + \text{sign}(X_3) + |X_4| + X_5 + X_5^2 - X_5^3 \tag{4.15}$$

3. Test Function 3:

$$y = f_1(X_1) + f_2(X_2) + f_3(X_3), \tag{4.16}$$

where $f_j = \beta_{j1}P_1(X_j) + \beta_{j2}P_2(X_j) + \beta_{j3}P_3(X_j),\ j = 1, 2, 3$ that consists of the Legendre polynomials of degree up to three and the standardised true coefficients are: $\boldsymbol{\beta} = (2, 1, 1/2, 1, 1, 1, -1, -4, 1)'_{9\times 1}$.

Test Function 1 is a linear function where BHS-NE is expected to benefit the most. We expect all methods except BHS-NE to do well in Test Function 2, where there is a highly nonlinear structure in the model. In Test Function 3, we expect HBGHS to perform well because the true functions are from basis functions.

Before testing our proposed methods, we first describe the procedure for generating true coefficients for the above three test functions. Suppose we are given a specified signal-to-noise ratio (SNR), a weighting vector $\mathbf{w} = (w_1, \ldots, w_m)$, the variance of true function α, and the unscaled coefficient $\boldsymbol{\beta}$. A weighting factor is generally a weight given to different data points to adjust their importance in a group. It allows us to give more or less importance to group members. In this paper, the weighting vector $\mathbf{w}$ was set to 1's in order to set every non-zero component to have equal variance. The procedure of generating true coefficients for above three functions is described as follows:

1. Generate $\mathbf{X}_{n\times m} = (\mathbf{X}_1, \ldots, \mathbf{X}_m)$ with a large sample size n from $U(-1, 1)$, where m is the number of non-zero components;

2. Now the unscaled function can be expressed as:

$$y' = \beta_1 f_1(X_1) + \cdots + \beta_m f_m(X_m);$$

3. Compute the scale for each component given the weighting vector $\mathbf{w}$ to ensure all components have same variance and contribute equally to the final function:

$$\alpha_1 = \sqrt{\frac{w_1}{\mathrm{Var}(\beta_1 f_1)}},\ \alpha_2 = \sqrt{\frac{w_2}{\mathrm{Var}(\beta_2 f_2)}},\ \cdots,\ \alpha_m = \sqrt{\frac{w_m}{\mathrm{Var}(\beta_m f_m)}},$$

and:

$$y^* = \alpha_1\beta_1 f_1(X_1) + \cdots + \alpha_m\beta_m f_m(X_m);$$

4. Compute the overall scale γ for the function to achieve α, the designated variance of y:

$$\gamma = \sqrt{\frac{\alpha}{\text{Var}(y^*)}},$$

the final function of y is

$$y = \gamma \left(\alpha_1\beta_1 f_1(X_1) + \cdots + \alpha_m\beta_m f_m(X_m)\right). \tag{4.17}$$

After specifying the true model, we apply our methods to a number of simulated data. The simulation procedure is:

1. Specify marginal functions for each dimension: $f_1, f_2, \cdots, f_m$;
2. If marginal functions $f_1, f_2, \cdots, f_m$ are splines functions or Legendre polynomials, apply the following procedure; otherwise, go to step 3;
   (a) For each column vector $\mathbf{X}_j$, $j = 1, \cdots, m$, we expand it using either spline functions at $K-1$ knots or Legendre polynomials of degree 2 to K, and then append it to $\mathbf{X}_j$ to obtain an expanded matrix $\tilde{\mathbf{X}}_j = (\tilde{\mathbf{X}}_{j1}, \cdots, \tilde{\mathbf{X}}_{jK})$, where $\tilde{\mathbf{X}}_{j1} = \mathbf{X}_j$;
   (b) Standardize each column $\mathbf{X}_j$, $j = 1, \cdots, m$, to have zero mean and unit length. Obtain the standardized matrix $\tilde{\mathbf{X}}_j$ and a scale vector $\mathbf{s}_j = (s_{j1}, \cdots, s_{jK})$;
   (c) Given a pre-specified standardized coefficient vector $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \cdots, \boldsymbol{\beta}_m)^T = (\beta_{11}, \cdots, \beta_{1K}, \cdots\cdots, \beta_{m,1}, \cdots, \beta_{m,K})^T$, calculate the true coefficient

$$\tilde{\boldsymbol{\beta}}_j = \left(\frac{\beta_{j1}}{s_{j1}}, \cdots, \frac{\beta_{jK}}{s_{jK}}\right)^T; \tag{4.18}$$

   (d) Each component is then calculated as: $f_j(\mathbf{X}_j) = \tilde{\mathbf{X}}_j\tilde{\boldsymbol{\beta}}_j$;
3. Compute the scale for each component to ensure all components have same variance:

$$\alpha_1^2 = \frac{\text{Var}(f_1)}{\text{Var}(f_1)},\ \alpha_2^2 = \frac{\text{Var}(f_1)}{\text{Var}(f_2)},\ \cdots,\ \alpha_m^2 = \frac{\text{Var}(f_1)}{\text{Var}(f_m)}; \tag{4.19}$$

4. Given SNR and the variance of the noise variable σ^2, compute

$$\gamma^2 = \frac{\text{SNR} \cdot \sigma^2}{\text{Var}(y^*)};$$

5. The final true function is:

$$y = \gamma[\alpha_1 f_1(\mathbf{X}_1) + \cdots + \alpha_m f_m(\mathbf{X}_m)]. \tag{4.20}$$

For all three test functions, we generated $p = 10$ predictors originally and expand it using different expansions on the original predictors, and varied the number of samples $n = \{100, 200\}$, the signal-to-noise ratio SNR$= \{1, 5, 10\}$ and the maximum degree of Legendre polynomial expansions $K = \{3, 6, 9, 12\}$. The detailed procedure is:

1. Generate $t = 100$ samples of $\mathbf{X}_{n\times p}$ from the uniform distribution $U(-1, 1)$;

2. For each realization, calculate $\mathbf{y}$ using the true function (4.17):

$$\mathbf{y} = \gamma\left(\alpha_1\beta_1 f_1(\mathbf{X}_1) + \cdots + \alpha_m\beta_m f_m(\mathbf{X}_m)\right) + \boldsymbol{\epsilon}, \tag{4.21}$$

   where m is the number of associated predictors, $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ and $\sigma^2 = \alpha/\text{SNR}$;

3. Expand $\mathbf{X}$ using Legendre polynomials with degree K and obtain $\tilde{\mathbf{X}}$;

4. Apply three methods to t samples of expanded data $\tilde{\mathbf{X}} = (\tilde{\mathbf{X}}_1, \ldots, \tilde{\mathbf{X}}_t)$: the regular Bayesian horseshoe method, the hierarchical Bayesian horseshoe method, and the lasso method with regularization parameter selected by the Bayesian information criterion (lasso-BIC);

5. Apply the regular Bayesian horseshoe method (BHS-NE) without expansions of the covariates $\mathbf{X}$ as our benchmark;

6. Obtain t posterior samples $\boldsymbol{\beta}$ for each method, and compute the mean $\boldsymbol{\beta}$ as the estimate.

4.4.3 Simulation Results

From the estimated coefficients, we calculated the mean squared error of the predictor using the following procedure:

1. Randomly generate $\mathbf{X}$ with a large sample size n from $U(-1, 1)$;
2. Calculate $\mathbf{y}$ from true functions using (4.17);
3. Expand $\mathbf{X}$ with Legendre polynomials: $\tilde{\mathbf{X}}$;
4. Calculate $\hat{\mathbf{y}}$ from the estimated coefficient for all three Bayesian models: $\hat{\mathbf{y}} = \tilde{\mathbf{X}}\hat{\boldsymbol{\beta}}$, where $\hat{\boldsymbol{\beta}}$ is the estimates from above three methods;
5. Compute the mean squared error of the predictor $\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$.

The BGHS method penalises variables only at group level and has very limited shrinkage performance in our designed test functions, especially as the degree of polynomial expansions increases. Therefore, the BGHS performed uniformly worse than the BHS and HBGHS in all experiments, and the results are not presented here.

The average mean squared prediction error and the corresponding standard deviation for Test Function 1 are shown in Table 4.1. In the simple linear test, the BHS model without expansion of covariates unsurprisingly consistently produced the smallest MSPE values. Of the remaining three methods, HBGHS and BHS were competitive for the sample size $n = 100$. As the sample size grew, HBGHS showed significant improvement in terms of MSPE compared to BHS, and the MSPE of HBGHS was smaller in most cases when $n = 200$. The prediction performance of BHS, HBGHS and lasso-BIC dropped as the degrees of expansion increased.

For Test Function 2, where there is a highly nonlinear relationship in the model, the HBGHS improved performance in almost all scenarios, as shown in Table 4.2. The only scenario where BHS slightly outperformed HBGHS was when $n = 100$ and the signal-to-noise ratio was low (SNR = 1). BHS-NE performed poorly in this case because it is unable to capture nonlinear effects.

TABLE 4.1: Mean and standard deviation (in parentheses) of squared prediction error for BHS, HBGHS, BHS-NE and lasso-BIC of Function 1, when sample size $n = \{100, 200\}$, signal-to-noise ratio SNR $= \{1, 5, 10\}$, and the highest degree of Legendre polynomial expansions $K = \{3, 6, 9, 12\}$.

n	SNR	Degree	BHS	HBGHS	BHS-NE	lasso-BIC
100	1	3	0.1137(0.0786)	0.1107(0.0861)	0.0856(0.0513)	0.1941(0.1090)
		6	0.1258(0.0918)	0.1322(0.1034)	0.0855(0.0520)	0.2543(0.1312)
		9	0.1536(0.1274)	0.1635(0.1423)	0.0852(0.0511)	0.5293(2.3280)
		12	0.2247(0.3075)	0.2200(0.2793)	0.0857(0.0519)	17.810(18.670)
	5	3	0.0178(0.0125)	0.0173(0.0125)	0.0154(0.0086)	0.0375(0.0190)
		6	0.0173(0.0115)	0.0176(0.0123)	0.0153(0.0088)	0.0491(0.0225)
		9	0.0194(0.0144)	0.0196(0.0143)	0.0153(0.0086)	0.1029(0.4565)
		12	0.0377(0.0797)	0.0346(0.0759)	0.0154(0.0087)	3.1820(3.6550)
	10	3	0.0088(0.0062)	0.0086(0.0062)	0.0077(0.0043)	0.0185(0.0090)
		6	0.0085(0.0056)	0.0087(0.0060)	0.0077(0.0044)	0.0251(0.0113)
		9	0.0095(0.0070)	0.0096(0.0070)	0.0077(0.0043)	0.0516(0.2269)
		12	0.0186(0.0394)	0.0171(0.0380)	0.0077(0.0044)	1.4160(1.7350)
200	1	3	0.0442(0.0258)	0.0422(0.0256)	0.0378(0.0200)	0.0954(0.0461)
		6	0.0461(0.0293)	0.0468(0.0294)	0.0377(0.0198)	0.1224(0.0457)
		9	0.0473(0.0316)	0.0467(0.0318)	0.0378(0.0202)	0.1354(0.0516)
		12	0.0497(0.0352)	0.0479(0.0317)	0.0376(0.0202)	0.1487(0.0544)
	5	3	0.0084(0.0051)	0.0082(0.0052)	0.0076(0.0040)	0.0194(0.0100)
		6	0.0087(0.0058)	0.0089(0.0059)	0.0076(0.0039)	0.0242(0.0100)
		9	0.0089(0.0062)	0.0088(0.0062)	0.0076(0.0040)	0.0269(0.0100)
		12	0.0094(0.0071)	0.0091(0.0063)	0.0076(0.0040)	0.0297(0.0109)
	10	3	0.0042(0.0026)	0.0041(0.0026)	0.0038(0.0020)	0.0094(0.0047)
		6	0.0043(0.0030)	0.0044(0.0030)	0.0038(0.0020)	0.0120(0.0044)
		9	0.0044(0.0031)	0.0044(0.0031)	0.0039(0.0020)	0.0133(0.0053)
		12	0.0047(0.0037)	0.0045(0.0032)	0.0038(0.0020)	0.0148(0.0056)

Marginal plots of Test Function 2 in 4.15 using HBGHS with $n = 200$, SNR= 5, Legendre polynomial expansions $K = 12$ are shown in Figure 4.2. The true marginal functions are indicated by a black solid line, the fitted models based on posterior mean, along with the $(25\%, 75\%)$ credible intervals are represented using dotted lines. The figure shows that the lasso-BIC fitted the true function poorly compared to both BHS and HBGHS because the bias in the lasso has caused lasso-BIC to underestimate the magnitude of the functions. Moreover, all three methods fitted marginal function 1, 2, 4 and 5 much better than they fitted marginal function 3 where the true function is a step function. Therefore, the Legendre polynomials are not an ideal approximation to the discrete function.

We now examine how Haar wavelets perform when the true function is a step

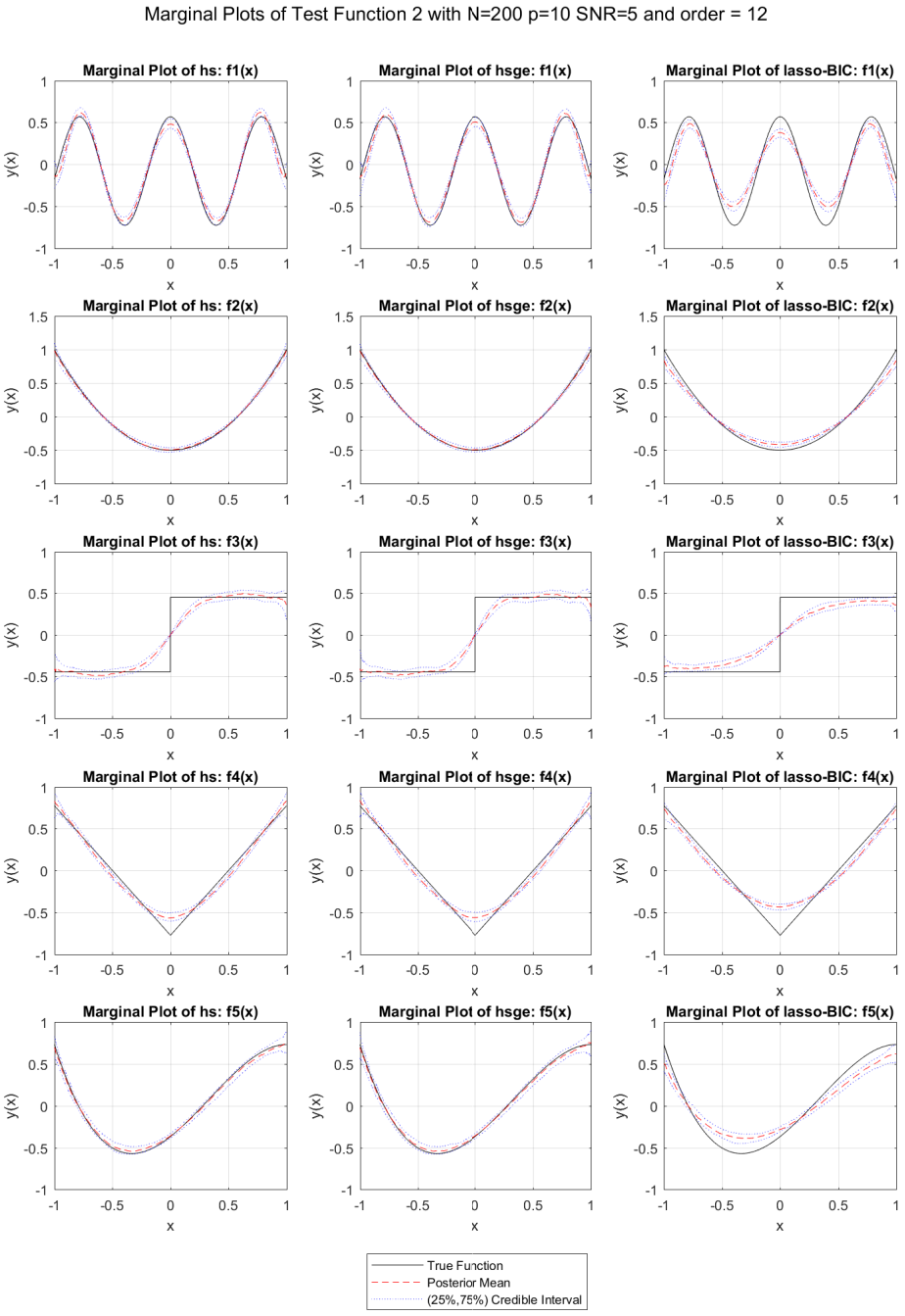

FIGURE 4.2: Marginal plots of Test Function 2 with sample size $n = 200$, SNR= 5, Legendre polynomial expansions $K = 12$ for BHS, HBGHS and lasso-BIC.

TABLE 4.2: Mean and standard deviation (in parentheses) of squared prediction error for BHS, HBGHS, BHS-NE and lasso-BIC of Function 2, when sample size $n = \{100, 200\}$, signal-to-noise ratio SNR $= \{1, 5, 10\}$, and the highest degree of Legendre polynomial expansions $K = \{3, 6, 9, 12\}$.

n	SNR	Degree	BHS	HBGHS	BHS-NE	lasso-BIC
100	1	3	0.5008(0.1204)	0.5122(0.1357)	0.9182(0.0756)	0.6512(0.2203)
		6	0.4968(0.1518)	0.5279(0.1721)	0.9208(0.0771)	0.6850(0.2236)
		9	0.5524(0.1757)	0.5816(0.1877)	0.9192(0.0757)	0.9774(1.8780)
		12	0.6591(0.2845)	0.6874(0.3629)	0.9199(0.0760)	18.150(20.220)
	5	3	0.3055(0.0382)	0.2947(0.0360)	0.8579(0.0488)	0.3523(0.0682)
		6	0.1728(0.0354)	0.1578(0.0321)	0.8593(0.0505)	0.2345(0.0656)
		9	0.1357(0.0495)	0.1246(0.0669)	0.8582(0.0497)	0.5216(0.9430)
		12	0.1882(0.1893)	0.1819(0.2279)	0.8582(0.0492)	3.4360(4.0510)
	10	3	0.2823(0.0293)	0.2739(0.0271)	0.8490(0.0415)	0.3189(0.0561)
		6	0.1363(0.0234)	0.1245(0.0206)	0.8501(0.0431)	0.1781(0.0430)
		9	0.0859(0.0381)	0.0792(0.0475)	0.8491(0.0423)	0.2571(0.4638)
		12	0.1287(0.1521)	0.1353(0.2316)	0.8488(0.0414)	1.6830(2.0180)
200	1	3	0.3306(0.0491)	0.3271(0.0485)	0.8452(0.0379)	0.3985(0.0765)
		6	0.2330(0.0560)	0.2220(0.0574)	0.8450(0.0372)	0.3417(0.1006)
		9	0.2125(0.0646)	0.1973(0.0661)	0.8453(0.0372)	0.3517(0.1147)
		12	0.2329(0.0783)	0.2185(0.0740)	0.8449(0.0372)	0.3820(0.1175)
	5	3	0.2545(0.0188)	0.2500(0.0179)	0.8170(0.0023)	0.2795(0.0275)
		6	0.1098(0.0154)	0.1051(0.0150)	0.8168(0.0231)	0.1464(0.0269)
		9	0.0602(0.0128)	0.0551(0.0133)	0.8172(0.0233)	0.1034(0.0289)
		12	0.0639(0.0178)	0.0580(0.0152)	0.8168(0.0231)	0.1168(0.0313)
	10	3	0.2442(0.0135)	0.2404(0.0128)	0.8136(0.0216)	0.2623(0.0187)
		6	0.0937(0.0109)	0.0906(0.0104)	0.8134(0.0215)	0.1157(0.0167)
		9	0.0406(0.0076)	0.0374(0.0085)	0.8138(0.0217)	0.0667(0.0150)
		12	0.0413(0.0104)	0.0373(0.0093)	0.8134(0.0215)	0.0721(0.0182)

function. Marginal plots of $\mathbf{X}_1$ and $\mathbf{X}_3$ for Test Function 2 using Legendre expansion of degree 12, Fourier expansion of degree 12 and Haar wavelet of degree 4 are shown in Figure 4.3. For marginal function f_1, both Legendre and Fourier expansion approximated the true function well. This is because they both consist of smooth basis functions, and so they approximated smooth function much better than the discrete basis functions did. However, for marginal function f_3 where the true function is a step function, the Legendre and Fourier obviously did not fit the true function as well as the Haar wavelet.

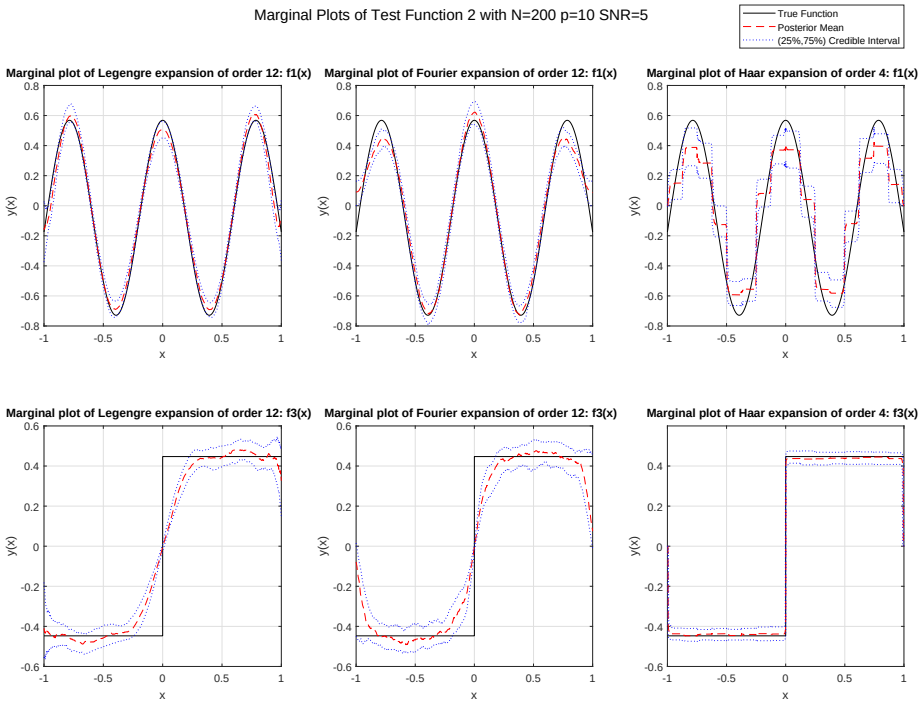

FIGURE 4.3: Marginal plots of Test Function 2 where $N = 200$, $p = 10$, SNR$= 5$ for Legendre expansions of degree 12, Fourier expansions of degree 12 and Haar wavelets of degree 4.

For Test Function 3, HBGHS was expected to achieve good performance because the true underlying additive model consists of a set of polynomial expansions. Indeed, referring to Table 4.3, HBGHS gave the smallest MSPE for all scenarios. The performance of the BHS is better than the lasso-BIC, and the BHS-NE unsurprisingly performed poorly for these polynomial test functions.

In general, the HBGHS method outperforms the regular BHS. As an illustration, boxplots of component-wise squared prediction errors for the BHS and the HBGHS methods when $p = 10$, $n = 100$, SNR $= 5$, $K = 3$ are shown in Figure 4.4. The first three components are associated with the response variable and the rest are noise variables. From the figure, we see that HBGHS penalises noise variables more than BHS. When all the variables within a group have small effects, HBGHS tends to apply extra shrinkage to the whole group and shrink all variables with the group simultaneously.

TABLE 4.3: Mean and standard deviation (in parentheses) of squared prediction error for BHS, HBGHS, BHS-NE and lasso-BIC of Function 3, when sample size $n = \{100, 200\}$, signal-to-noise ratio SNR $= \{1, 5, 10\}$, and the highest degree of Legendre polynomial expansions $K = \{3, 6, 9, 12\}$.

n	SNR	Degree	BHS	HBGHS	BHS-NE	lasso-BIC
100	1	3	0.1564(0.0705)	0.1336(0.0699)	0.5514(0.0678)	0.2398(0.1146)
		6	0.1809(0.0789)	0.1645(0.0824)	0.5521(0.0677)	0.3128(0.1434)
		9	0.2142(0.0989)	0.2019(0.1079)	0.5514(0.0669)	0.6967(2.5440)
		12	0.2876(0.3553)	0.2808(0.4541)	0.5520(0.0670)	18.090(20.070)
	5	3	0.0329(0.0158)	0.0262(0.0141)	0.4951(0.0260)	0.0555(0.0247)
		6	0.0366(0.0172)	0.0318(0.0163)	0.4953(0.0261)	0.0732(0.0308)
		9	0.0421(0.0197)	0.0394(0.0194)	0.4951(0.0260)	0.1852(0.5761)
		12	0.0647(0.0969)	0.0618(0.1343)	0.4954(0.0256)	3.3320(4.3170)
	10	3	0.0179(0.0084)	0.0139(0.0073)	0.4896(0.0211)	0.0287(0.0127)
		6	0.0201(0.0086)	0.0173(0.0081)	0.4899(0.0212)	0.0386(0.0143)
		9	0.0230(0.0105)	0.0215(0.0105)	0.4896(0.0211)	0.0977(0.2842)
		12	0.0357(0.0521)	0.0339(0.0674)	0.4899(0.0209)	1.4920(1.9740)
200	1	3	0.0756(0.0318)	0.0629(0.0297)	0.4995(0.0259)	0.1235(0.0495)
		6	0.0905(0.0374)	0.0809(0.0355)	0.4994(0.0258)	0.1709(0.0702)
		9	0.0993(0.0413)	0.0894(0.0374)	0.4996(0.0262)	0.1918(0.0735)
		12	0.1056(0.0432)	0.0955(0.0396)	0.4993(0.0258)	0.2089(0.0763)
	5	3	0.0166(0.0069)	0.0133(0.0064)	0.4786(0.0110)	0.0289(0.0113)
		6	0.0192(0.0082)	0.0173(0.0080)	0.4787(0.0110)	0.0382(0.0125)
		9	0.0206(0.0087)	0.0189(0.0080)	0.4787(0.0110)	0.0456(0.0165)
		12	0.0218(0.0094)	0.0201(0.0082)	0.4785(0.0110)	0.0502(0.0178)
	10	3	0.0091(0.0038)	0.0071(0.0034)	0.4763(0.0093)	0.0150(0.0060)
		6	0.0105(0.0048)	0.0095(0.0045)	0.4763(0.0094)	0.0209(0.0073)
		9	0.0112(0.0050)	0.0104(0.0044)	0.4763(0.0093)	0.0242(0.0081)
		12	0.0118(0.0056)	0.0111(0.0047)	0.4762(0.0093)	0.0266(0.0084)

4.5 Real Data

To evaluate the performance of the grouped Bayesian horseshoe methods, we apply them to three real data sets: the diabetes data, the Boston housing data and the electrical-maintenance data.

To perform the real data analysis, we use hold-out validation methods by dividing data into two subsets, where the training data contains 75% of the samples and the testing data contains 25% of the samples. We fit the Bayesian group horseshoe model to the training data with predictor $\mathbf{X}$ expanded using three expansions: Legendre, Fourier and Haar, and then compute the mean of posterior samples as

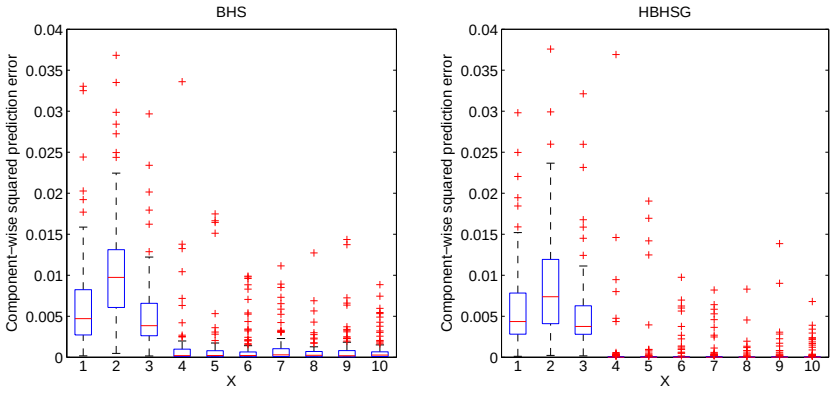

FIGURE 4.4: Boxplots of component-wise prediction errors for the BHS and the HBGHS when there are $p = 10$ predictors, $n = 100$ samples, SNR $= 5$ and $K = 3$ degree of Legendre polynomial expansions. The first three components are associated with the response variable.

our estimate. To compare the performance for all models, we compute the mean squared prediction error and its standard deviation.

4.5.1 Diabetes Data

For the diabetes data set, there are $n = 442$ samples and $p = 10$ predictors. The second variable X_2 is a categorical variable with only two levels (0 and 1) representing the sex and hence it directly enters the model as a predictor. The remaining variables are continuous variables and they are expanded using three different expansions: Legendre polynomials, Fourier, Haar wavelets with various degrees $K = \{1, 2, \ldots, 8\}$ before running the tests.

The mean squared prediction error along with its standard deviation for Legendre expansions is shown in Table 4.4. The Legendre expansion with degree one is the original model where all predictors are not expanded. From the table, we see that the BHS model with Legendre polynomials expansions of degree 3 produced the smallest prediction error compared to all other models across different degrees. As the degree of expansions grows from degree 3, the predictive performance became worse for all models (BHS, BGHS, HBGHS, lasso-BIC) and the performance deteriorated.

TABLE 4.4: Mean and standard deviation (in parentheses) of squared prediction error for BHS, BGHS, HBGHS and lasso-BIC using Legendre expansions with degree from 1 to 8 for diabetes data.

Degree	BHS	BGHS	HBGHS	lasso-BIC
1	3038.52(298.35)	3026.20(306.93)	3040.15(299.27)	3058.97(295.20)
2	3014.52(295.17)	3026.59(301.28)	3041.54(293.18)	3101.72(294.31)
3	2982.44(304.25)	3019.45(318.91)	3015.82(301.02)	3089.81(307.32)
4	3020.62(296.98)	3100.55(324.74)	3055.53(291.05)	3119.49(295.67)
5	3060.85(295.14)	3222.33(578.98)	3093.03(292.72)	3149.09(306.17)
6	3096.13(315.43)	3427.45(1230.24)	3142.54(335.82)	3176.23(312.50)
7	5845.92(24269.50)	5768.45(13105.63)	4974.10(17883.82)	5118.33(19233.38)
8	7588.42(23030.02)	19911.14(80035.00)	7716.27(24382.87)	5997.58(21029.81)

The mean squared prediction error along with its standard deviation for all models using Fourier expansions is shown in Table 4.5. Order zero corresponds to non expansions of the predictor. When the data is not expanded, BGHS produced the smallest mean prediction error compared to BHS, HBGHS and lasso-BIC. For expanded data, the BHS model with one degree of expansion produced the smallest mean prediction error compared to all other models. Compared to Table 4.4, the Fourier expansions showed robustness as degree of expansions increased. The Fourier expansions do not break down because the mean squared prediction error did not explode as the order of expansion increased, and therefore, are very stable.

TABLE 4.5: Mean and standard deviation (in parentheses) of squared prediction error for BHS, BGHS, HBGHS and lasso-BIC using Fourier expansions with degree from 0 to 8 for diabetes data.

Degree	BHS	BGHS	HBGHS	lasso-BIC
0	3038.52(298.35)	3026.2(306.93)	3040.15(299.27)	3058.97(295.2)
1	2994.95(314.33)	3015.56(316.50)	3028.58(311.26)	3086.42(298.98)
2	3065.95(308.46)	3099.64(323.23)	3097.34(306.94)	3125.72(306.76)
3	3083.45(308.93)	3139.96(344.21)	3114.96(307.96)	3161.08(300.83)
4	3111.24(298.9)	3156.81(351.14)	3141.37(303.22)	3173.41(311.84)
5	3136.42(300.99)	3193.53(335.03)	3163.42(301.64)	3201.53(310.79)
6	3153.86(298.41)	3234.12(336.35)	3181.61(298.77)	3216.43(313.75)
7	3170.17(296.12)	3226.73(328.3)	3201.76(301.87)	3224.88(314.58)
8	3180.85(298.70)	3232.33(336.51)	3214.17(309.72)	3237.43(313.68)

The mean squared prediction error along with its standard deviation for all models using Haar expansions is shown in Table 4.6. From the table, the degree of zero

corresponds to the BHS-NE method, and it always gave the smallest values in terms of mean squared prediction error for all four methods. Although the Haar wavelet expansions for BHS, BGHS, HBGHS or lasso-BIC do not work well for the diabetes data compared to the BHS-NE, its performance did not deteriorate until degree 7 for BGHS and degree 6 for lasso-BIC respectively. It is noticed here that a Haar expansion of degree 8 actually consist of 2^8 basis functions which is equivalent to high degree in both Legendre or Fourier expansions. BGHS did not work as well as the BHS or HBGHS for higher degree because BGHS does not perform individual shrinkage within each group, and therefore, it is worse especially for the case where there exist a small number of strong effects in a large group. In this example, BGHS performed worse for large degrees of expansions.

TABLE 4.6: Mean and standard deviation (in parentheses) of squared prediction error for BHS, BGHS, HBGHS and lasso-BIC using Haar expansions with degree from 0 to 8 for diabetes data.

Degree	BHS	BGHS	HBGHS	lasso-BIC
0	3038.52(298.35)	3026.2(306.93)	3040.15(299.27)	3058.97(295.2)
1	3048.27(299.03)	3034.4(304.48)	3072.34(300.39)	3109.4(299.21)
2	3078.08(292.37)	3086.32(303.79)	3115.9(288.37)	3168.8(296.74)
3	3086.37(291.43)	3132.54(325.99)	3106.43(293.23)	3144.95(302.04)
4	3161.63(295.94)	3240.81(358.11)	3169.79(301.33)	3209.46(318.43)
5	3228.01(314.03)	3481.7(390.81)	3239.31(316.32)	3232.43(319.44)
6	3263.13(317.12)	3736.69(439.69)	3277.82(320.13)	13321.46(1912.49)
7	3285(315.07)	4171.02(510.07)	3286.48(316.28)	8482.69(887.34)
8	3287.95(315.28)	4398.19(527.41)	3294.56(326.14)	3444.17(681.77)

4.5.2 Boston Housing Data

For the Boston housing data, there are $n = 506$ samples and $p = 13$ predictors where X_4 and X_9 are categorical variables. Since covariate X_4 only contains two levels (0 and 1), it enters the model directly. The covariate X_9 consist of 9 levels and therefore can be expressed as 8 dummy variables. The remaining variables are continuous variables and are expanded using Legendre polynomials, Fourier expansions and Haar wavelets.

The mean squared prediction error along with its standard deviation for Legendre expansions is shown in Table 4.7. It is obvious that Legendre expansion of degree 1 for BGHS model returned the smallest mean prediction error for all methods with all degrees of expansions. For all methods with degree less than 4, the mean prediction error for BHS-NE was no better than all Bayesian-type models with expansions. Therefore, there exists non-linear relationship in predictors for Boston housing data. However, when degree increased to 4, all methods including BHS, BGHS, HBGHS started to break down while the lasso-BIC performed better than the Bayesian models. The Legendre expansion is not very stable because the mean spared prediction errors increased substantially as degree expanded.

TABLE 4.7: Mean and standard deviation (in parentheses) of squared prediction error for BHS, BGHS, HBGHS and lasso-BIC using Legendre expansions with degree from 1 to 8 for Boston housing data.

Degree	BHS	BGHS	HBGHS	lasso-BIC
1	23.85(5.29)	23.76(5.43)	24.07(5.39)	23.97(5.47)
2	17.48(5.41)	17.35(5.35)	17.73(5.43)	18.20(5.63)
3	17.60(5.61)	17.93(6.05)	18.19(5.66)	18.41(5.64)
4	34.39(130.98)	42.09(175.05)	50.91(236.16)	17.38(5.36)
5	161.24(1248.33)	123.94(752.97)	202.94(1542.70)	51.86(334.89)
6	40.30(132.87)	43.85(138.10)	48.97(218.25)	25.38(49.64)
7	6114.56(46937.15)	81224.09(572865.63)	25353.98(193151.77)	90.67(145.27)
8	4010.61(34499.45)	19984.00(190718.21)	8490.64(76360.66)	27.98(35.41)

For Boston housing data with Fourier expansions, the mean squared prediction error along with its standard deviation is shown in Table 4.8. From the table, we see that BGHS consistently produced the smallest prediction errors for all degrees of expansions compared to all other models (BHS, HBGHS, lasso-BIC). BHS-NE gave the largest values in terms of mean squared prediction error. Therefore, there is non-linear effects in the data, which agrees with results from the Legendre expansion.

For the Haar wavelet expansion, the mean squared prediction error along with its standard deviation is shown in Table 4.9. It is clear that BGHS with degree 3 produced the smallest prediction error across all degrees and all methods in the table. The BHS-NE, which is any method with degree 0, produced the largest

TABLE 4.8: Mean and standard deviation (in parentheses) of squared prediction error for BHS, BGHS, HBGHS and lasso-BIC using Fourier expansions with degree from 0 to 8 for Boston housing data.

Degree	BHS	BGHS	HBGHS	lasso-BIC
0	23.85(5.29)	23.76(5.43)	24.07(5.39)	23.97(5.47)
1	17.22(5.40)	17.02(5.30)	17.48(5.45)	18.14(5.71)
2	17.19(5.23)	16.36(5.02)	17.54(5.24)	18.63(5.64)
3	16.54(4.89)	15.57(4.5)	17.19(5.19)	17.74(5.14)
4	15.95(4.53)	14.35(3.94)	16.54(5.15)	18.09(5.18)
5	16.10(4.48)	14.55(3.96)	16.69(4.82)	18.43(5.02)
6	15.84(4.38)	14.24(3.87)	16.37(4.48)	18.22(5.19)
7	15.97(4.3)	14.53(3.94)	16.51(4.67)	17.82(5.33)
8	16.36(4.64)	14.94(4.20)	16.80(4.71)	18.19(5.33)

mean prediction error compared to all four methods with expansions. In general, the Haar wavelet method is stable because the mean prediction error did not break down until degree reaches 8 for BHS and HBGHS, and 6 for BGHS. As we mentioned, the Haar wavelet expansion of degree 8 consists of 2^8 basis functions, which is equivalent to large value in both Legendre or Fourier expansions.

TABLE 4.9: Mean and standard deviation (in parentheses) of squared prediction error for BHS, BGHS, HBGHS and lasso-BIC using Haar expansions with degree from 0 to 8 for Boston housing data.

Degree	BHS	BGHS	HBGHS	lasso-BIC
0	23.85(5.29)	23.76(5.43)	24.07(5.39)	23.97(5.47)
1	23.87(5.34)	23.72(5.17)	24.38(5.25)	24.54(5.44)
2	19.39(5.89)	19.47(5.72)	19.68(5.91)	20.09(5.68)
3	17.38(5.05)	16.27(4.74)	17.79(5.09)	17.96(4.90)
4	18.61(7.42)	16.83(5.40)	19.71(7.91)	18.57(5.35)
5	17.53(5.74)	17.12(7.14)	19.42(7.86)	18.09(6.07)
6	18.89(7.99)	20.83(7.71)	19.88(10.42)	17.30(5.15)
7	19.65(8.41)	26.52(7.35)	18.57(8.12)	17.23(4.89)
8	26.53(11.62)	29.74(7.5)	24.34(7.85)	17.65(4.90)

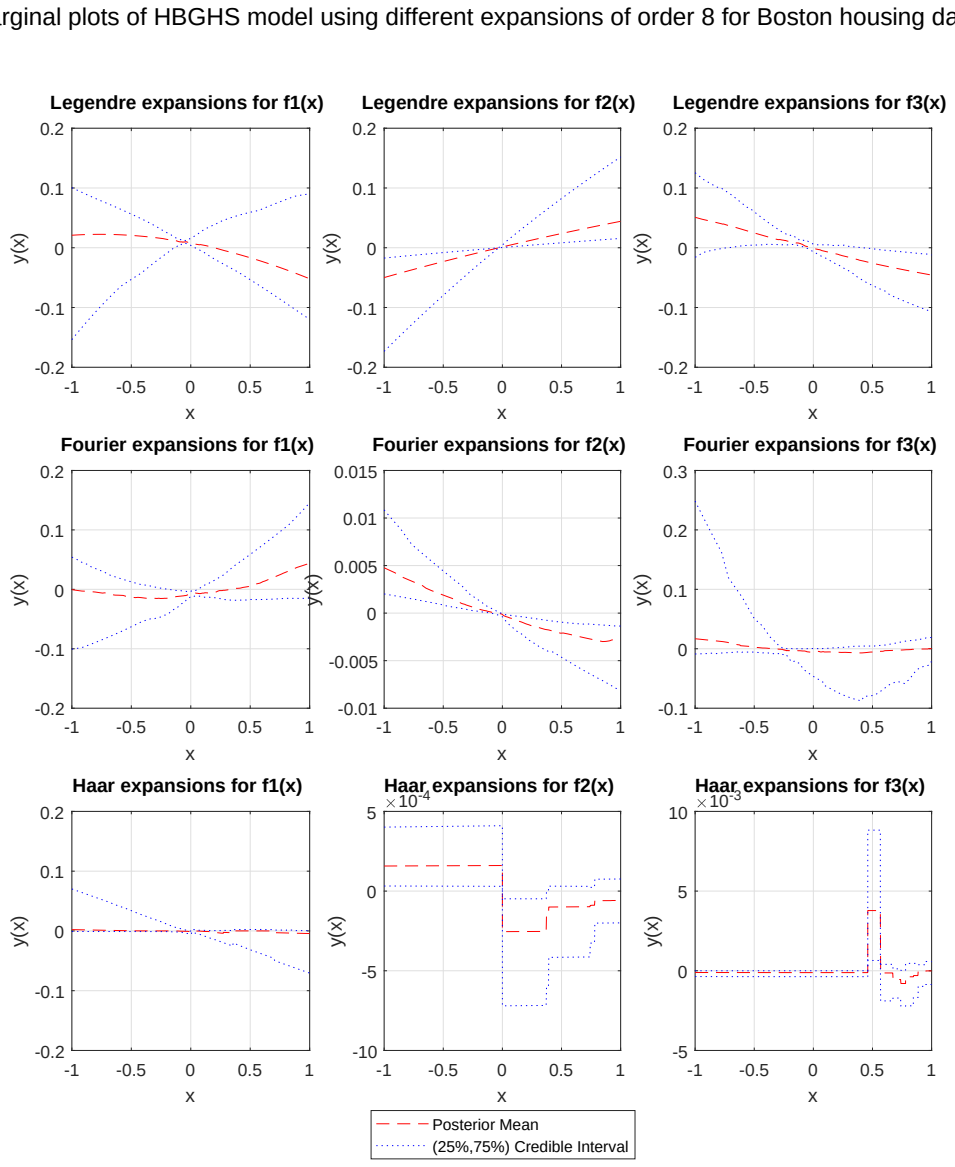

FIGURE 4.5: Marginal plots of Boston housing data with three expansions: Legendre polynomial, Fourier, Haar wavelet, all of degree $K = 8$ for the HBGHS model.

Figure 4.5 shows marginal plots for Legendre, Fourier and Haar wavelet expansions with degree $K = 8$ based on HBGHS model.

4.5.3 Electrical-Maintenance Data

To evaluate the performance of the BGHS method and the HBGHS method, we applied them to the Electrical-Maintenance data set [144]. There are $n = 1056$ samples and $p = 4$ input variables in the Electrical-Maintenance data set. All variables are continuous.

All predictors were expanded using Legendre polynomials with degrees $K = \{2, 4, 6, 8, 10\}$ for each of the following methods: BHS, BGHS, HBGHS and lasso-BIC. We also tested BHS-NE with original non-expanded predictors as the benchmark.

The mean and standard deviation (in parentheses) of mean squared prediction errors are shown in Table 4.10. We see that there clearly exists a nonlinear pattern in the data because the mean squared prediction error decreases as the degree of expansions increases. On average, HBGHS has the lowest MSPE, especially when the degree of polynomials grows. BGHS has the smallest MSPE when each group has few variables ($K = 2$).

TABLE 4.10: Mean and standard deviation of squared prediction errors for Electrical-Maintenance data set.

Degree	BHS	BGHS	HBGHS	BHS-NE	lasso-BIC
2	26866.9(2636)	26855.0(2637)	26879.8(2637)	27309.7(2489)	26858.6(2632)
4	13405.1(1285)	13437.1(1285)	13394.0(1281)	27314.3(2493)	13488.9(1291)
6	12939.2(1257)	13061.4(1255)	12939.9(1254)	27312.0(2498)	13038.6(1252)
8	11019.9(1141)	11054.0(1097)	10978.4(1136)	27307.9(2492)	11067.9(1100)
10	9970.9(1096)	10057.1(1075)	9958.4(1097)	27316.6(2492)	10022.4(1079)

4.6 Conclusion

In conclusion, we have proposed the BGHS method in Section 4.2.2 and the HBGHS method in Section 4.2.3 for performing both group-wise and within group selection.

We applied our proposed HBGHS model to a set of test functions and compared with the regular BHS in terms of the mean squared prediction error on both simulated data (see Section 4.4).

The proposed methods frequently outperform the regular BHS method when applied to nonlinear functions and additive models, particularly when there are variables that are associated with the target. The reason that our proposed methods outperform the regular BHS method is that they are capable of shrinking whole groups of basis expansions that are associated with that variable to zero if there is not enough evidence of association. This property is particularly important in performing group variable selections because when there is not enough evidence in group association, all variables in that group are expected to be penalised more heavily. Even when there is no underlying group structure, the proposed methods are competitive with the regular BHS method. The results of real data analysis also support their promising performance (see Section 4.5).

Part of the work in this chapter has been published as:

Z. Xu, D. F. Schmidt, E. Makalic, G. Qian, and J. L. Hopper, "Bayesian grouped horseshoe regression with application to additive models," in *Australasian Joint Conference on Artificial Intelligence*, pp. 229–240, Springer, 2016.

Chapter 5

Bayesian Sparse Global-Local Shrinkage Regression for Selection of Grouped Variables

5.1 Introduction

Consider the standard Gaussian linear regression model

$$\mathbf{y} = \mu\mathbf{1} + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

where $\mathbf{y} = (y_1, \ldots, y_n)^T$ are observations of a response variable, $\mathbf{X} = (\mathbf{x}_1^T, \ldots, \mathbf{x}_n^T)^T$ is an $n \times p$ observation matrix of the predictor variables, μ is the intercept, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ is a vector of regression coefficients to be estimated from the data and $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)^T$ are independent Gaussian random disturbances.

In recent years, estimation of high-dimensional regression models for big data has become a prominent problem in many fields of research. Here, the number of regression coefficient parameters to be estimated, p, is assumed to be of high dimension and often exceeds the number of observations, n. A further assumption when analysing high-dimensional data by a regression model is that most of the

predictors are unassociated with the response variable. In other words, the vector $\boldsymbol{\beta}$ is assumed to be sparse.

Conventional approaches to estimating regression coefficients, such as least-squares and the method of maximum likelihood, may break down and perform poorly when applied to high-dimensional data. Penalised regression methods, especially those based on ℓ_1 regularisation, such as the Lasso [9], provide popular alternatives to tackle these difficulties (see Section 2.3.4). As an example, the Lasso is able to perform both shrinkage and variable selection by minimising a goodness of fit term plus a penalty proportional to the sum of the absolute values of the regression coefficients estimates. This may force some of the coefficients to be estimated as exactly zero; therefore the Lasso can produce sparse estimates.

Most penalised regression methods can be viewed as Bayesian procedures by interpreting the estimate of $\boldsymbol{\beta}$ as the posterior mode under an appropriate prior distribution. For example, the Lasso procedure can be viewed as maximum a posteriori estimation under a Gaussian linear regression model when the coefficients $\boldsymbol{\beta}$ follow independent and identical double exponential prior distributions; this is known as the Bayesian Lasso [59] (see Section 2.4.2.2).

Two broad classes of priors are commonly used in Bayesian sparse estimation and variable selection (see Section 2.4): (1) two-component finite mixture priors, and (2) continuous shrinkage priors. A two-component finite mixture prior, also known as a spike-and-slab prior [55, 145], consists of a point mass at $\beta_j = 0$ and an absolute continuous alternative density for $\beta_j \neq 0$. Although the two-component finite mixture prior can produce estimates of regression coefficients exactly equal to zero, it entails an exploration of a model space of size 2^p; computational difficulties therefore arise as the number of predictors p grows.

For the continuous shrinkage prior approach, each β_j is assigned a continuous shrinkage prior centered at $\beta_j = 0$. One of the most important classes of continuous shrinkage priors are the global-local shrinkage priors [12] where the prior

distribution for $\boldsymbol{\beta}$ can be expressed using the hierarchy

$$\beta_j \,|\, \lambda_j^2, \tau^2 \;\sim\; N(0, \tau^2\lambda_j^2), \quad \lambda_j^2 \;\sim\; \pi(\lambda_j^2), \quad \tau^2 \;\sim\; \pi(\tau^2).$$

Here, λ_j is the local shrinkage parameter associated with the jth predictor, τ is the global shrinkage parameter that controls the overall degree of shrinkage, and $\pi(\cdot)$ are generic prior distributions or density functions. The particular choice of the prior distribution for the local shrinkage parameters $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_p)$ determines the behaviour of the resulting Bayesian estimator. This global-local hierarchy includes many important models such as the horseshoe [62], the horseshoe+ [63], the Dirichlet–Laplace [13], etc.

This continuous shrinkage approach offers great computational advantages in comparison with the two-component spike-and-slab mixture priors: there is no requirement to explore a discrete model space of size 2^p. However, one limitation of the continuous shrinkage prior approach is that it fails to provide a sparse solution. To address the problem, specialised methods for producing sparse estimates from posterior samples, such as penalised variable selection based on posterior credible regions [68] or the decoupled shrinkage and selection method [14], have been developed (see Section 2.4.2.8).

Without loss of generality, we assume that the response variable is centered and the covariates are standardised to have a zero mean and unit variance. A general hierarchical Bayesian global-local shrinkage regression model can be expressed as

$$\begin{aligned}
\mathbf{y} \,|\, \mathbf{X}, \boldsymbol{\beta}, \sigma^2 \;&\sim\; N(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n) \\
\boldsymbol{\beta} \,|\, \sigma^2, \tau^2, \lambda_1^2, \ldots, \lambda_p^2 \;&\sim\; N(\mathbf{0}, \sigma^2\tau^2\mathbf{D}_{\boldsymbol{\lambda}}) \\
\mathbf{D}_{\boldsymbol{\lambda}} &= \operatorname{diag}(\lambda_1^2, \ldots, \lambda_p^2) \\
\lambda_j^2 \;&\sim\; \pi(\lambda_j^2) d\lambda_j^2, \quad j = 1, \ldots, p \\
\tau \;&\sim\; C^+(0, 1) \\
\sigma^2 \;&\sim\; \sigma^{-2} d\sigma^2,
\end{aligned} \tag{5.1}$$

where τ is the global shrinkage parameter, λ_j are the local shrinkage parameters and $C^+(0,1)$ denotes the standard half-Cauchy distribution. Commonly used prior distributions for the local shrinkage parameters, λ_j, include

$$\begin{aligned} \lambda_j^2 &\sim \text{Exp}(1), \quad j = 1, \ldots, p \quad \text{for the Lasso,} \\ \lambda_j &\sim C^+(0,1), \quad j = 1, \ldots, p \quad \text{for the horseshoe,} \\ \lambda_j \sim C^+(0,\phi_j),\ \phi_j &\sim C^+(0,1), \quad j = 1, \ldots, p \quad \text{for the horseshoe+,} \end{aligned} \tag{5.2}$$

respectively, where Exp(1) denotes the standard exponential distribution.

5.1.1 Selection of Variable Groups

Although possessing many desirable theoretical, computational, and empirical properties, continuous shrinkage techniques so far have focused primarily on identifying important individual predictors. In many practical situations, it is however of great interest to select groups of variables, or find sparse combinations of grouped variables. This can be seen in additive models where each continuous predictor is expressed as a linear combination of basis functions such that selection of important predictors becomes equivalent to selection of groups (linear combinations) of the basis functions (see Section 4.1.1). This can also be seen in genome-wide association studies, in which the aim is to determine which single nucleotide polymorphisms are associated with a disease or trait. Since the single nucleotide polymorphisms can be conceptually grouped into different genes, and the genes can be grouped into different "pathways", it is of great interest to be able to select important groups of single nucleotide polymorphisms.

Under the assumption that a group structure is given, the aim is to find sparsity at both the group level, as well as within the groups themselves. An important group selection procedure is the group Lasso [37] that solves the convex optimisation problem

$$\hat{\boldsymbol{\beta}}(\kappa) = \arg\min_{\boldsymbol{\beta}} \left\{ \|\mathbf{y} - \sum_{g=1}^{G} \mathbf{X}_g \boldsymbol{\beta}_g\|_2^2 + \kappa \sum_{g=1}^{G} \sqrt{s_g} \|\boldsymbol{\beta}_g\|_2 \right\},$$

where $\mathbf{X}_g$ is the submatrix of $\mathbf{X}$ containing those predictors in group g, $\boldsymbol{\beta}_g$ is the coefficient vector for group g, s_g is the number of predictors in group g, $\|\cdot\|_2$ is the Euclidean norm, and $\boldsymbol{\beta} = (\boldsymbol{\beta}_1', \ldots, \boldsymbol{\beta}_G')'$. A potential weakness of the group Lasso is that it only provides a sparse solution for a set of groups, and does not provide sparsity at an individual predictor level. To select individual predictors as well as groups of predictors, the sparse-group Lasso [10] was proposed

$$\hat{\boldsymbol{\beta}}(\kappa_1, \kappa_2) = \arg\min_{\boldsymbol{\beta}} \left\{ \|\mathbf{y} - \sum_{g=1}^{G} \mathbf{X}_g \boldsymbol{\beta}_g\|_2^2 + \kappa_1 \sum_{g=1}^{G} \|\boldsymbol{\beta}_g\|_2 + \kappa_2 \|\boldsymbol{\beta}\|_1 \right\},$$

which introduces ℓ_1-type penalisation at both variable and group levels; when $\kappa_2 = 0$, the sparse-group Lasso reduces to the group Lasso, and when $\kappa_1 = 0$ it reduces to the regular Lasso. The Lasso-based approaches described above, as well as extensions of these methods, require the estimation of tuning parameters. Cross-validation methods [146, 147] are frequently used to meet this requirement; however, selection of tuning parameters by cross-validation may fail to achieve consistent variable selection [148]. A particular weakness of all non-Bayesian sparse estimation procedures is the difficulty in obtaining confidence intervals for the resulting estimates [149].

The Bayesian formulation of the sparse group regression model works by specifying a hierarchical or multi-level prior distribution over the regression parameters that incorporates prior beliefs about the group structure of the predictors. The specific hierarchy ensures that the hyper-parameters can be automatically estimated by the inherent Bayesian procedure. Additionally, as the usual approach to analysing Bayesian models is through Monte-Carlo Markov chain (MCMC) methods, posterior samples can be used to easily obtain credible intervals for all parameters.

Recently, a Bayesian sparse group selection model (BSGS) which is based on two-component mixture priors is proposed [77]. Under the sparsity assumption, only a small number of groups are active in explaining the response variable, and only a small number of predictors are active within each of the active groups. The BSGS model consists of two layers of indicator variables: the group indicators

which determine whether each group is active or not, and the individual indicators which determine whether each particular predictor is active or not. The BSGS uses a group-wise Gibbs sampler for the posterior sampling, and has demonstrated superior performance in selecting active groups, as well as identifying the active predictors within selected groups, in comparison to the sparse group Lasso. However, just as with the non-group based spike-and-slab methods, a problem with the two-component finite mixture model approach is the requirement for exploration of an exponentially growing model space, which can result in problems of computational intractability.

5.1.2 Decoupled Shrinkage and Selection Methods

Bayesian regression estimators based on continuous shrinkage priors have their own limitation in that they do not automatically yield sparse solutions. Even though shrinkage priors such as the horseshoe do promote sparsity, the usual posterior mean or median estimates of the regression coefficients never equal zero almost surely. Sparse models, in which some of the regression coefficients equal zero exactly, are of great value since they are more interpretable, particularly when p is large.

Therefore, it is highly desirable to develop some Bayesian variable selection techniques that produce sparse solutions from non-zero posterior samples of regression coefficients. Progress has already been made in this regard; for example, Bondell and Reich [68], and Zhang and Bondell [150] use variable selection techniques based on credible regions to convert posterior samples to sparse estimates; Hahn and Carvalho [14] use conventional penalised regression estimation to produce sparse estimates directly from the posterior samples, and name their method the decoupled shrinkage and selection (DSS) procedure.

The posterior mean frequently exhibits excellent performance in terms of prediction [151]; the key idea underlying the DSS is to determine a regression coefficient

vector that is sparse while remaining close in some sense to its posterior mean estimate, and therefore retaining good predictive performance. The DSS procedure works by forming a new target vector

$$\bar{\mathbf{y}} = \mathbf{X}\bar{\boldsymbol{\beta}}, \tag{5.3}$$

where $\bar{\boldsymbol{\beta}} = \mathbb{E}\left[\boldsymbol{\beta} \,|\, \mathbf{y}\right]$ is the posterior mean of the regression coefficients; the vector $\bar{\mathbf{y}}$ is a smoothed version of the original data vector $\mathbf{y}$ that incorporates the effects of the shrinkage induced by the prior distribution on the coefficients $\boldsymbol{\beta}$. From this, a DSS loss function $L(\cdot)$ can be defined

$$L(\boldsymbol{\gamma}) = \kappa\|\boldsymbol{\gamma}\|_0 + n^{-1}\|\bar{\mathbf{y}} - \mathbf{X}\boldsymbol{\gamma}\|_2^2, \tag{5.4}$$

where $\|\boldsymbol{\gamma}\|_0 = \sum_{j=1}^{p} \mathbb{1}(\gamma_j \neq 0)$ is the ℓ_0-norm and $\kappa > 0$ is a penalisation parameter. This loss function is a combination of a goodness-of-fit term that measures the closeness of the regression coefficient $\boldsymbol{\gamma}$ to the posterior mean $\bar{\boldsymbol{\beta}}$, plus a penalty term that measures the "complexity" of the model, and induces sparsity. The coefficient vector

$$\boldsymbol{\beta}_\kappa = \underset{\boldsymbol{\gamma}}{\operatorname{argmin}} \left\{\kappa\|\boldsymbol{\gamma}\|_0 + n^{-1}\|\mathbf{X}\bar{\boldsymbol{\beta}} - \mathbf{X}\boldsymbol{\gamma}\|_2^2\right\}. \tag{5.5}$$

then represents a particular sparsification of $\bar{\boldsymbol{\beta}}$, with the degree of sparsification determined by the choice of κ. For $\kappa = 0$, no sparsification occurs; for larger values of κ, the vector $\boldsymbol{\beta}_\kappa$ becomes increasingly sparse.

The optimal solution is achieved through the trade-off between the number of variables in the model and the predictive performance in terms of the mean-squared prediction errors. If a variable X_j has a strong association with the response variable, the posterior mean, $\bar{\beta}_j$, associated with this variable will tend to be large, and therefore, X_j is more likely to be included in the final sparsified model. Conversely, if a variable X_j has little or no effect on the response variable, it is more likely to be removed from the final model because $\bar{\beta}_j$ will be small. Therefore, the DSS algorithm builds connections between the Bayesian methods and the

popular penalised likelihood approaches, and produces sparse estimates with good predictive performance.

For a given value of κ, the solution of (5.5), $\boldsymbol{\beta}_\kappa$, is a potentially sparse coefficient vector. To use the DSS method, an obvious question is how to choose an appropriate value of κ. In the original DSS paper [14], two statistics were proposed to help visualise the predictive deterioration that quantifies the goodness of $\boldsymbol{\beta}_\kappa$ in predicting $\boldsymbol{\beta}$ due to sparsification for varying values of κ:

1. the variation explained by a sparsified linear predictor

$$\rho^2_\kappa = \frac{n^{-1}\|\mathbf{X}\boldsymbol{\beta}\|^2}{n^{-1}\|\mathbf{X}\boldsymbol{\beta}\|^2 + \sigma^2 + n^{-1}\|\mathbf{X}\boldsymbol{\beta} - \mathbf{X}\boldsymbol{\beta}_\kappa\|^2}; \tag{5.6}$$

2. the excess error of a sparsified linear predictor

$$\psi_\kappa = \sqrt{n^{-1}\|\mathbf{X}\boldsymbol{\beta} - \mathbf{X}\boldsymbol{\beta}_\kappa\|^2 + \sigma^2} - \sigma. \tag{5.7}$$

The smallest model for which the 90% ρ^2_κ credible interval contains $\mathbb{E}\left[\rho^2_0 \,|\, \mathbf{y}\right]$, where ρ^2_0 is the variation explained by the posterior mean, is recommended as a good sparse approximation of $\bar{\boldsymbol{\beta}}$. Although this approach produces a single model, it is heuristic in nature, and there is no compelling reason to select a 90% credible interval, in place of say, a 95% interval.

5.1.3 Our Contribution

In this chapter, we extend the Bayesian model with continuous global-local shrinkage priors to handle group structures that include overlapping and multilevel group structures; in particular, we extend the recent horseshoe and horseshoe+ priors for grouped predictor structures. After obtaining the posterior samples from the Bayesian models, we extend the decoupled shrinkage and selection method for grouped variables and apply the group non-negative garrotte to produce sparsified estimates. Generalised information criteria using posterior expected estimates of

the degrees of freedom of the models in the group DSS model path are then used to select the final model, and perform group-wise selection.

5.2 Grouped Global-Local Shrinkage Models

In regression, predictors with similar characteristics can often be collected into groups. For example, in bioinformatics, genetic variants used to predict the risk of a disease, can be formed into groups corresponding to particular genes. Generally, only a small number of genes are assumed to have important associations with the disease and we wish to select these important genes (i.e., perform sparse group selection). In addition, there exist many alternative groupings of genetic variants such as pathways, groups consisting of variants with similar function, etc. We propose to model such problems using a Bayesian multi-level hierarchy which can capture alternative schemes for grouping predictors (see Figure 5.1).

We assume that each level of our multi-level hierarchy consists of non-overlapping groups. Two groups are said to be non-overlapping if they share no predictors in common. Assuming that groups within a level of our hierarchy are mutually exclusive greatly simplifies the sampling procedure. However, between different levels of the hierarchy, we allow the specification of overlapping groups which enables modelling of arbitrary grouping structures.

Consider a multi-level hierarchy consisting of $K + 2$ levels. Within this multi-level hierarchy, we define a local level consisting of p groups each containing a single predictor and a global level which consists of a single group containing all p predictors. Levels 1 through K contain non-overlapping groupings of predictors determined by the application at hand. This approach generalises the global-local shrinkage hierarchy discussed in Section 5.1. The local shrinkage parameters $(\lambda_1, \ldots, \lambda_p)$ that control the amount of shrinkage for individual predictors are located at the first level of the hierarchy. At the last level of the hierarchy is the global shrinkage parameter τ which determines the overall level of shrinkage for all predictors. To perform estimation and selection for grouped variables within a

global-local shrinkage hierarchy, we introduce group shrinkage parameters associated with each group of predictors. We denote the group shrinkage parameter for the g-th group at the k-th level of the hierarchy by $\delta_{k,g}$. These group shrinkage parameters induce an additional level of shrinkage on all predictors in a group.

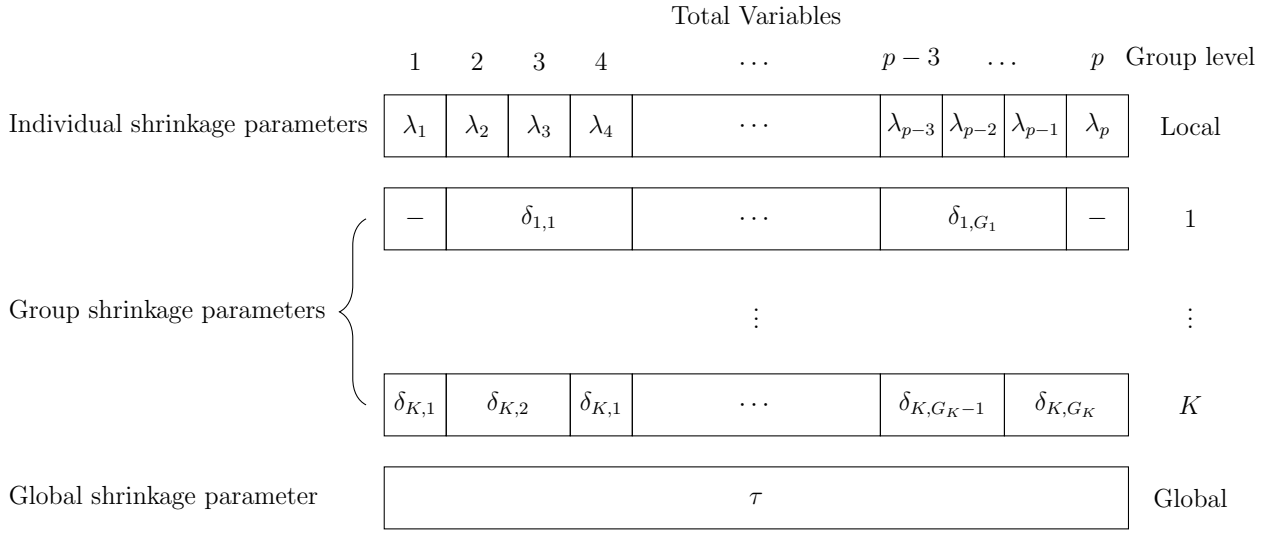

FIGURE 5.1: An illustration of possible group structures of p variables with one level of individual variables, K levels of grouped variables and one level of total variables.

Figure 5.1 shows an example of a possible multilevel group structure. Suppose there are p variables and $K+2$ levels of specified groups of predictors. At the first group level, each of the p predictors belongs to a separate group of size one. In the last level, all predictors are grouped together into one group of size p. Figure 5.1 clearly demonstrates the flexibility of our proposed multi-level approach. There is no requirement that each predictor be assigned to a group at any of the K levels outside the local and the global level; for example, in level 1 variables 1 and p are not assigned to a group. In addition, the groups are not required to be contiguous and can contain any combination of the p predictors; for example, group 1 at level K contains both predictor 1 and predictor 4.

5.2.1 Bayesian Grouped Global-Local Shrinkage Hierarchical Models

Formally, let $G(k,j) \in \{0,\ldots,G_k\}$ denote the group containing predictor j at level k of our hierarchy where $G_k \geq 1$ is the number of groups defined at level k. If predictor j does not belong to any group at level k we set $G(k,j) = 0$. The hierarchical representation of our proposed Bayesian grouped global-local shrinkage regression model can then be written as

$$\begin{aligned}
\mathbf{y} \,|\, \mathbf{X}, \boldsymbol{\beta}, \sigma^2 &\sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I}_n) \\
\boldsymbol{\beta} \,|\, \sigma^2, \tau^2, \lambda_j, \delta_{k,g} &\sim \mathcal{N}(\mathbf{0}, \sigma^2\tau^2 \mathbf{D}_{\boldsymbol{\lambda}} \mathbf{D}_{\boldsymbol{\delta}_1} \cdots \mathbf{D}_{\boldsymbol{\delta}_K}), \\
\lambda_j &\sim \pi(\lambda_j) d\lambda_j, && j = 1,\ldots,p \\
\delta_{k,g} &\sim \pi(\delta_{k,g}) d\delta_{k,g}, && g = 1,\ldots,G_k,\ k = 1,\ldots,K \\
\tau &\sim C^+(0,1) \\
\sigma^2 &\sim \sigma^{-2} d\sigma^2
\end{aligned} \tag{5.8}$$

where

$$\begin{aligned}
\mathbf{D}_{\boldsymbol{\lambda}} &= \operatorname{diag}(\lambda_1^2,\ldots,\lambda_p^2) \\
\mathbf{D}_{\boldsymbol{\delta}_k} &= \operatorname{diag}(\Omega_{k,1},\ldots,\Omega_{k,p}), && k = 1,\ldots,K \\
\Omega_{k,j} &= \mathbb{1}\{\omega_{k,j} = 0\} + \omega_{k,j} \cdot \mathbb{1}\{\omega_{k,j} \neq 0\}, && j = 1,\ldots,p,\ k = 1,\ldots,K \\
\omega_{k,j} &= \sum_{g=1}^{G_k} \delta_{k,g}^2 \cdot \mathbb{1}\{G(k,j) = g\}, && j = 1,\ldots,p,\ k = 1,\ldots,K
\end{aligned} \tag{5.9}$$

Here, $\boldsymbol{\lambda} = (\lambda_1,\ldots,\lambda_p)$ are the local shrinkage parameters associated with the individual predictors and $\boldsymbol{\delta}_1 = (\delta_{1,1},\ldots,\delta_{1,G_1}),\ldots,\boldsymbol{\delta}_K = (\delta_{K,1},\ldots,\delta_{K,G_K})$ are the group shrinkage parameters.

The global shrinkage parameter, τ, controls the overall degree of shrinkage of the coefficients while the local shrinkage parameters, λ_j, determine the shrinkage for each individual coefficient. The grouped shrinkage parameters $\delta_{k,j}$ provide additional shrinkage for predictors within groups. In the special case where $K = 0$

(i.e., there are only a local and a global levels), the model reduces to the usual Bayesian global-local shrinkage regression model (5.1).

This model offers more flexibility than the regular Bayesian model (5.1) by shrinking individual predictors separately, and allowing shrinkage for groups of predictors simultaneously. The ability to specify complex group structures using multiple levels further increases the flexibility of the proposed model. In many real-world applications, each predictor may belong to more than one group. For example, in statistical genomics, predictors corresponding to single nucleotide polymorphisms may belong to different genes, and a collection of genes may form a biological pathway. Additionally, pathways may share genes and genes may share single nucleotide polymorphisms. By utilising an appropriate number of grouping levels, the complexity of this situation is easily handled by our proposed hierarchical model.

After the model is specified, posterior samples of $\boldsymbol{\beta}$, σ^2, τ, λ_j, $\delta_{k,g}$ can be obtained from the full conditional distributions through the implementation of a Gibbs sampler, which cyclically updates one parameter at a time given the current values of all the other parameters and hyper-parameters. To sample the global hyper-parameter τ, we use the sampler proposed by [67] which decomposes

$$\tau \sim C^{+}(0, a) \tag{5.10}$$

into

$$\tau^2 \,|\nu \sim \mathcal{IG}\left(\frac{1}{2}, \frac{1}{\nu}\right) \text{ and } \nu \sim \mathcal{IG}\left(\frac{1}{2}, \frac{1}{a^2}\right), \tag{5.11}$$

where $\mathcal{IG}$ denotes the inverse gamma distribution.

The posterior distributions for $\boldsymbol{\beta}$, σ^2 and τ do not depend on the choice of prior distributions for the local shrinkage parameters λ_j and group shrinkage parameters $\delta_{k,g}$. Therefore, the full conditional distributions of $\boldsymbol{\beta}$, σ^2 and τ are the same for the Bayesian group Lasso, the Bayesian group horseshoe and the Bayesian group

horseshoe+ models. The full conditional distribution of $\boldsymbol{\beta}$ is

$$\begin{aligned}\boldsymbol{\beta}\,|\cdot &\sim \mathcal{N}\left(\mathbf{A}^{-1}\mathbf{X}^T\mathbf{y}, \sigma^2\mathbf{A}^{-1}\right)\\ \mathbf{A} &= \mathbf{X}^T\mathbf{X} + (\tau^2\mathbf{D}_{\boldsymbol{\lambda}}\mathbf{D}_{\boldsymbol{\delta}_1}\cdots\mathbf{D}_{\boldsymbol{\delta}_K})^{-1}.\end{aligned}$$

The full conditional distribution of σ^2 is

$$\sigma^2\,|\cdot \sim \mathcal{IG}\left(\frac{n-1+p}{2}, \frac{(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^T(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) + \boldsymbol{\beta}^T(\tau^2\mathbf{D}_{\boldsymbol{\lambda}}\mathbf{D}_{\boldsymbol{\delta}_1}\cdots\mathbf{D}_{\boldsymbol{\delta}_K})^{-1}\boldsymbol{\beta}}{2}\right).$$

The full conditional distributions of τ and ν are

$$\begin{aligned}\tau^2\,|\cdot &\sim \mathcal{IG}\left(\frac{p+1}{2}, \frac{\boldsymbol{\beta}^T(\mathbf{D}_{\boldsymbol{\lambda}}\mathbf{D}_{\boldsymbol{\delta}_1}\cdots\mathbf{D}_{\boldsymbol{\delta}_K})^{-1}\boldsymbol{\beta}}{2\sigma^2} + \frac{1}{\nu}\right)\\ \nu\,|\cdot &\sim \mathcal{IG}\left(1, \frac{1}{\tau^2}+1\right),\end{aligned}$$

where sampling from $\mathcal{IG}\,(1, 1/\tau^2+1)$ is equivalent to sampling from an exponential distribution. The full conditional distributions of λ_j and $\delta_{k,g}$ depend on the particular choice of the prior distribution and are discussed in the next section.

5.2.2 Bayesian Group Lasso Model

Park and Casella [59] proposed the Bayesian Lasso in which each regression coefficient follows an independent double-exponential (Laplace) distribution. This idea was extended to the Bayesian group Lasso [152], which we now further extend to our multilevel group hierarchy. The Laplace prior of the Lasso model can be expressed as a Gaussian variance mixture distribution where the mixing density is an exponential distribution. This implies that the Bayesian Lasso can be written as a Bayesian global-local shrinkage hierarchy where the prior distributions for the local shrinkage parameters λ_j are

$$\lambda_j^2 \sim \text{Exp}(1),\ j = 1, \ldots, p. \tag{5.12}$$

Let $\boldsymbol{\delta}$ denote the collection of $\delta_{k,g}$, the full conditional distribution of λ_j^{-2}, $j = 1, \ldots, p$ is then

$$\lambda_j^{-2} \,|\, \boldsymbol{\beta}, \sigma^2, \tau^2, \boldsymbol{\delta} \sim \text{InvGaussian}\left(\alpha_j^{\frac{1}{2}}, 2\right),$$

where

$$\alpha_j = \frac{2\sigma^2\tau^2[\mathbf{D}]_{j,j}}{\beta_j^2} \quad \text{and} \quad \mathbf{D} = \mathbf{D}_{\boldsymbol{\delta}_1} \cdots \mathbf{D}_{\boldsymbol{\delta}_K}.$$

Here $[\mathbf{D}]_{j,j}$ is the entry in the jth row and jth column of the matrix $\mathbf{D}$. Now we choose the prior distribution for the group shrinkage parameter $\delta_{k,g}^2$ associated with group g at level k to be

$$\delta_{k,g}^2 \sim \text{Exp}(1), \; g = 1, \ldots, G_k, \; k = 1, \ldots, K, \tag{5.13}$$

where G_k is the number of predictor groups at level k of our hierarchy. The full conditional distribution of $\delta_{k,g}^{-2}$ is then a generalised inverse Gaussian (GIG) distribution of the form

$$\delta_{k,g}^{-2} \,|\, \boldsymbol{\beta}, \sigma^2, \tau^2, \boldsymbol{\lambda} \sim \text{GIG}\left(\alpha_{k,g}, 2, \frac{s_{k,g}}{2} - 1\right),$$

where

$$\alpha_{k,g} = \frac{1}{\sigma^2\tau^2} \sum_{i \in l(k,g)} \frac{\beta_i^2}{\lambda_i^2[\mathbf{D}_{-k}]_{i,i}},$$

$s_{k,g}$ is the number of predictors in group g at level k and $l(k,g) = \{j \in \{1, \ldots, p\} : G(k,j) = g\}$ is the set of predictors at level k that belongs to group g. Here

$$\mathbf{D}_{-k} = \prod_{i=1, i \neq k}^{K} \mathbf{D}_{\boldsymbol{\delta}_i} \tag{5.14}$$

where $\mathbf{D}_{\boldsymbol{\delta}_i}$ is given by (5.9).

5.2.3 Bayesian Group Horseshoe Model

The horseshoe prior has been demonstrated to have good performance in high dimensional and sparse regression problems, leaving strong signals untouched and

aggressively shrinking noise variables [65]. The prior distribution for the local shrinkage parameters, λ_j, in the horseshoe model is

$$\lambda_j \sim C^+(0,1), \ j = 1, \ldots, p, \tag{5.15}$$

where $C^+(0,1)$ is the standard half-Cauchy distribution. By introducing auxiliary variables in similar fashion to (5.10) and (5.11), the standard half-Cauchy distribution can be expressed as

$$\lambda_j^2 \,|\, c_j \sim \mathcal{IG}\left(\frac{1}{2}, \frac{1}{c_j}\right), \ c_j \sim \mathcal{IG}\left(\frac{1}{2}, 1\right).$$

The full conditional distributions for $\lambda_1^2, \ldots, \lambda_p^2$ and the auxilary variables $c_1, \ldots, c_p$ are

$$\lambda_j^2 \,|\, \boldsymbol{\beta}, \sigma^2, \tau^2, \boldsymbol{\delta}, c_j \sim \mathcal{IG}\left(1, \frac{1}{\alpha_j} + \frac{1}{c_j}\right), \quad \alpha_j = \frac{2\sigma^2\tau^2[\mathbf{D}]_{j,j}}{\beta_j^2}$$
$$c_j \,|\, \lambda_j^2 \sim \mathcal{IG}\left(1, \frac{1}{\lambda_j^2} + 1\right).$$

The group shrinkage parameters $\delta_{k,g}^2$ in the group horseshoe also follow the standard half-Cauchy distribution [153]

$$\delta_{k,g} \sim C^+(0,1), \ g = 1, \ldots, G_k, \ k = 1, \ldots, K\,. \tag{5.16}$$

Using the half-Cauchy decomposition from (5.10) and (5.11), these prior distributions can also be written as

$$\delta_{k,g}^2 \,|\, t_{k,g} \sim \mathcal{IG}\left(\frac{1}{2}, \frac{1}{t_{k,g}}\right), \ t_{k,g} \sim \mathcal{IG}\left(\frac{1}{2}, 1\right). \tag{5.17}$$

The full conditional distributions for $\delta_{k,g}^2$ and $t_{k,g}$ are then

$$\delta_{k,g}^2 \,|\, \boldsymbol{\beta}, \sigma^2, \tau^2, \boldsymbol{\lambda}, t_{k,g} \sim \mathcal{IG}\left(\frac{s_{k,g}+1}{2}, \frac{1}{2\sigma^2\tau^2} \sum_{i \in l(k,g)} \frac{\beta_i^2}{\lambda_i^2[\mathbf{D}_{-k}]_{i,i}} + \frac{1}{t_{k,g}}\right)$$
$$t_{k,g} \,|\, \delta_{k,g}^2 \sim \mathcal{IG}\left(1, \frac{1}{\delta_{k,g}^2} + 1\right),$$

where $s_{k,g}$ is the number of predictors in group g at level k, $l(k,g) = \{j \in \{1,\ldots,p\} : G(k,j) = g\}$ is the set of predictors at level k that belongs to group g and $\mathbf{D}_{-k}$ is given by (5.14).

5.2.4 Bayesian Group Horseshoe+ Model

The horseshoe+ prior has a heavier tail than the regular horseshoe prior, resulting in more aggressive shrinkage of weak signals and producing sparser estimates than the regular horseshoe prior. The prior distribution of the local shrinkage parameters, λ_j, for the horseshoe+ model is

$$\lambda_j \,|\, \phi_j \;\sim\; C^{+}(0,\phi_j),\;\; \phi_j \;\sim\; C^{+}(0,1),\;\; j = 1,\ldots,p. \tag{5.18}$$

By using the decomposition in (5.10) and (5.11), the above prior can be written as

$$\lambda_j^2 \,|\, c_j \;\sim\; \mathcal{IG}\left(\frac{1}{2},\frac{1}{c_j}\right),\; c_j \,|\, \phi_j^2 \;\sim\; \mathcal{IG}\left(\frac{1}{2},\frac{1}{\phi_j^2}\right),\; \phi_j^2 \,|\, \eta_j \;\sim\; \mathcal{IG}\left(\frac{1}{2},\frac{1}{\eta_j}\right),\; \eta_j \;\sim\; \mathcal{IG}\left(\frac{1}{2},1\right).$$

The full conditional distributions for λ_j^2 and associated auxilary variables are

$$\begin{aligned}
\lambda_j^2 \,|\, \boldsymbol{\beta},\sigma^2,\tau^2,\boldsymbol{\delta},c_j \;&\sim\; \mathcal{IG}\left(1,\frac{\beta_j^2}{\alpha_j}+\frac{1}{c_j}\right), \quad \alpha_j = 2\sigma^2\tau^2[\mathbf{D}]_{j,j} \\
c_j \,|\, \lambda_j^2,\phi_j^2 \;&\sim\; \mathcal{IG}\left(1,\frac{1}{\lambda_j^2}+\frac{1}{\phi_j^2}\right) \\
\phi_j^2 \,|\, c_j,\eta_j \;&\sim\; \mathcal{IG}\left(1,\frac{1}{c_j}+\frac{1}{\eta_j}\right) \\
\eta_j \,|\, \phi_j^2 \;&\sim\; \mathcal{IG}\left(1,\frac{1}{\phi_j^2}+1\right).
\end{aligned}$$

The horseshoe+ prior distribution for the group shrinkage parameters $\delta_{k,g}^2$ is

$$\delta_{k,g} \,|\, t_{k,g} \;\sim\; C^{+}(0,t_{k,g}),\;\; t_{k,g} \;\sim\; C^{+}(0,1).$$

By using the decomposition, the above prior can be written as

$$\delta_{k,g}^2 \,|\, \xi_{k,g} \sim \mathcal{IG}\left(\frac{1}{2}, \frac{1}{\xi_{k,g}}\right), \; \xi_{k,g} \,|\, t_{k,g}^2 \sim \mathcal{IG}\left(\frac{1}{2}, \frac{1}{t_{k,g}^2}\right)$$
$$t_{k,g}^2 \,|\, \psi_{k,g} \sim \mathcal{IG}\left(\frac{1}{2}, \frac{1}{\psi_{k,g}}\right), \; \psi_{k,g} \sim \mathcal{IG}\left(\frac{1}{2}, 1\right).$$

Using the decomposition (5.10) and (5.11), the full conditional distributions of $\delta_{k,g}^2$ and the associated auxiliary variables are

$$\begin{aligned}
\delta_{k,g}^2 \,|\, \boldsymbol{\beta}, \sigma^2, \tau^2, \boldsymbol{\lambda}, t_{k,g} &\sim \mathcal{IG}\left(\frac{s_{k,g}+1}{2}, \frac{1}{2\sigma^2\tau^2} \sum_{i \in l(k,g)} \frac{\beta_i^2}{\lambda_i^2 [\mathbf{D}_{-k}]_{i,i}} + \frac{1}{\xi_{k,g}}\right) \\
\xi_{k,g} \,|\, \delta_{k,g}^2, t_{k,g}^2 &\sim \mathcal{IG}\left(1, \frac{1}{\delta_{k,g}^2} + \frac{1}{t_{k,g}^2}\right) \\
t_{k,g}^2 \,|\, \xi_{k,g}, \psi_{k,g} &\sim \mathcal{IG}\left(1, \frac{1}{\xi_{k,g}} + \frac{1}{\psi_{k,g}}\right) \\
\psi_{k,g} \,|\, t_{k,g}^2 &\sim \mathcal{IG}\left(1, \frac{1}{t_{k,g}^2} + 1\right).
\end{aligned}$$

where $s_{k,g}$ is the number of predictors in group g at level k, $l(k,g) = \{j \in \{1,\ldots,p\} : G(k,j) = g\}$ is the set of predictors at level k that belongs to group g, and $\mathbf{D}_{-k}$ is given by (5.14).

5.3 Decoupled Shrinkage and Selection for Grouped Variables

Recall that in the DSS procedure, the sparsification of $\bar{\boldsymbol{\beta}}$ involves solving a penalised regression problem using the smoothed data vector $\bar{\mathbf{y}} = \mathbf{X}\bar{\boldsymbol{\beta}}$:

$$\boldsymbol{\beta}_\kappa = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left\{ n^{-1} \|\bar{\mathbf{y}} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \kappa \|\boldsymbol{\beta}\|_0 \right\}. \tag{5.19}$$

There are two potential limitations of the original DSS procedure: (i) it does not generalise to the problem of selecting groups of variables, and (ii) it does not provide an objective and general procedure for selecting which of the models in

the DSS solution path should be used as the final, sparsified estimate. In this section we propose an extension of the DSS procedure, referred to as group DSS, which extends the original DSS proposal to accommodate selection of groups, and discuss how to adapt the standard information criteria based approaches to model selection to the DSS framework.

5.3.1 Group Decoupled Shrinkage and Selection Methods

The original DSS method produces sparse estimates at the level of individual variables rather than groups. Consider the multi-level group hierarchy discussed in Section 5.2. If we restrict interest to sparse estimation at a single level of the hierarchy, the ideas underlying DSS can easily be extended to deal with grouped variables, as none of the groups at a given level overlap. For the purposes of the group DSS procedure, variables that do not belong to any group are considered to form their own singleton groups. To select sparse estimates for grouped variables, we propose to use the solution to the following ideal group DSS optimisation problem:

$$\boldsymbol{\beta}_\kappa = \underset{\boldsymbol{\beta}_1,\ldots,\boldsymbol{\beta}_G}{\operatorname{argmin}} \left\{ n^{-1}\|\mathbf{X}\bar{\boldsymbol{\beta}} - \sum_{g=1}^{G} \mathbf{X}_g\boldsymbol{\beta}_g\|_2^2 + \kappa \sum_{g=1}^{G} s_g \mathbb{1}(\|\boldsymbol{\beta}_g\|_0 \neq 0) \right\}, \tag{5.20}$$

where $\mathbf{X}_g$ is the submatrix of $\mathbf{X}$ containing all predictors in group g, $\boldsymbol{\beta}_g$ are the regression coefficients associated with group g, s_g is the number of predictors in group g and G is the total number of groups at the level of the hierarchy we are interested in.

However, one potential problem in solving the optimisation problem (5.20) is that the counting penalty $\|\boldsymbol{\beta}_g\|_0$ results in an intractable optimisation problem when the number of groups is moderate to large. A standard approach to address this issue is to approximate the ℓ_0 penalty with an ℓ_1 penalty; for example, in the case of the original DSS algorithm (5.5), Hahn and Carvalho [14] proposed the

following surrogate ℓ_1 optimisation problem

$$\boldsymbol{\beta}_\kappa = \underset{\boldsymbol{\gamma}}{\operatorname{argmin}} \left\{ \kappa \sum_{j=1}^{p} \frac{|\gamma_j|}{|w_j|} + n^{-1} \|\mathbf{X}\bar{\boldsymbol{\beta}} - \mathbf{X}\boldsymbol{\gamma}\|_2^2 \right\},$$

where $w_j = \bar{\beta}_j$ is the posterior mean for coefficient j. This approach is analogous to the adaptive Lasso [11] in which w_j is replaced by the least-squares estimate $\hat{\beta}_j$, and can be easily solved by a rescaling of the design matrix $\mathbf{X}$.

A similar approach may be used to extend DSS to grouped variables. For example, we could replace the ℓ_0-norm in (5.20) by the grouped Lasso penalty [37]

$$\frac{1}{2}\|\mathbf{y} - \sum_{g=1}^{G} \mathbf{X}_g \boldsymbol{\beta}_g\|_2^2 + \kappa \sum_{g=1}^{G} \|\boldsymbol{\beta}_g\|_{\mathbf{K}_g},$$

where $\|\boldsymbol{\beta}_g\|_{\mathbf{K}_g} = (\boldsymbol{\beta}_g^T \mathbf{K}_g \boldsymbol{\beta}_g)^{1/2}$ and $\mathbf{K}_g$ are some symmetric positive definite matrices. However, a difficulty in using the group Lasso is that the corresponding solution path is no longer piecewise linear and is potentially slow to compute.

In this paper, we use the group non-negative garrotte (NNG) in place of the group Lasso. We adopted this procedure as it is relatively simple to use, and its formulation is similar in nature to the adaptive Lasso. Furthermore, the group NNG reduces the dimensionality of the solution path to the number of groups. An approximate solution to (5.20) can be achieved through the group non-negative garrotte by solving

$$\mathbf{d}_\kappa = \underset{\mathbf{d}}{\operatorname{argmin}} \left(\frac{1}{2}\|\bar{\mathbf{y}} - \mathbf{Z}\mathbf{d}\|^2 + \kappa \sum_{g=1}^{G} s_g d_g \right), \quad \text{subject to } d_g \geq 0, \forall g, \tag{5.21}$$

where $\bar{\mathbf{y}} = \mathbf{X}\bar{\boldsymbol{\beta}}$, $\mathbf{Z} = (\mathbf{Z}_1, \ldots, \mathbf{Z}_G)$, $\mathbf{Z}_g = \mathbf{X}_g \bar{\boldsymbol{\beta}}_g$, s_g is the size of group g and $\mathbf{d} = (d_1, \ldots, d_G)$ is a vector of group shrinkage coefficients. The algorithm for computing the non-negative garotte solution path is similar to the group least angle regression algorithm, and the solution to the group DSS for a given κ can be computed as $\boldsymbol{\beta}_\kappa = (\bar{\boldsymbol{\beta}}_1[\mathbf{d}_\kappa]_1, \ldots, \bar{\boldsymbol{\beta}}_G[\mathbf{d}_\kappa]_G)$.

The group DSS described above only works on a single level of group structure. However, it can be easily extended to multilevel group structures; for example, after obtaining posterior samples, group DSS can be run separately at each level to select groups of variables for every level.

5.3.2 Information Criteria for Group DSS

The group DSS procedure produces a number of candidate sparse models by varying the regularisation parameter κ in (5.21), but does not specify which model should be chosen. To assist in selecting a sparse model, the original DSS procedure recommended a heuristic selection criterion based on credible intervals of two statistics: (i) the variation explained (5.6), and (ii) the excess error (5.7). Although both of these heuristics are easily adapted to grouped variables, the corresponding selection criteria still require the arbitrary choice of a credible interval size. We now propose an alternative approach to selecting κ by adapting the standard information criteria approach commonly used in the model selection literature.

A standard approach for selecting a sparse linear model from a set of candidate models is through the use of generalised information criteria (GIC). The usual formulation of the generalised information criteria consists of a likelihood function that measures the goodness-of-fit of a model under consideration, and a model complexity penalty term that is a function of the degrees of freedom of the model; namely

$$\text{GIC}(\boldsymbol{\beta}, \sigma^2) = L(\boldsymbol{\beta}, \sigma^2) + \alpha(k, n, \sigma^2) \tag{5.22}$$

in the context of linear regression model selection, where $L(\cdot)$ is the negative log-likelihood function, $\boldsymbol{\beta}$ are the model coefficients, σ^2 is the unknown noise variance, $\alpha(\cdot)$ is the penalty function, n is the sample size and $k = ||\boldsymbol{\beta}||_\infty$ is the degrees of freedom of the model. We then select the model that attains the minimum GIC value over the set of candidate models we are considering.

There exists a wide range of information criteria in the statistical model selection literature; one of the most well-known information criterion is the Bayesian information criterion (BIC) [48], which is given by

$$\mathrm{BIC}(\boldsymbol{\beta}, \sigma^2) = L(\boldsymbol{\beta}, \sigma^2) + \frac{k}{2}\ln(n), \tag{5.23}$$

where n is the number of observations. The BIC is derived as an asymptotic approximation of the marginal probability of the data. Alternatives to BIC, that utilise more accurate approximations at the cost of stronger assumptions, include information theoretic minimum message length (MML) [154] and minimum description length [155] principles. A recent MML criterion for linear-Gaussian regression models [53] is

$$\mathrm{MML}_u(\boldsymbol{\beta}, \sigma^2) = L(\boldsymbol{\beta}, \sigma^2) + \left(\frac{k+1}{2}\right)\ln\left(\frac{\|\mathbf{y}\|^2}{2\sigma^2}\right) - \Gamma\left(\frac{k+3}{2}\right) + \frac{1}{2}\ln\left(k+1\right), \tag{5.24}$$

which has been shown to perform well in a wide range of settings [156]. In contrast to BIC, the MML_u penalty term is a function of the signal-to-noise ratio of the fitted model.

Another widely used information criterion is the Akaike information criterion (AIC) [47]

$$\mathrm{AIC}(\boldsymbol{\beta}, \sigma^2) = L(\boldsymbol{\beta}, \sigma^2) + k. \tag{5.25}$$

The AIC penalises the number of parameters less strongly than the BIC, and in the case when the sample size n is small, or the degrees of freedom k is large relative to n, the AIC tends to select models with too many parameters, and therefore has an increased probability of overfitting. To address this issue, Hurvich and Tsai [51] proposed the corrected AIC

$$\mathrm{AIC}_c(\boldsymbol{\beta}, \sigma^2) = L(\boldsymbol{\beta}, \sigma^2) + \frac{kn}{n-k-1}. \tag{5.26}$$

Details regarding derivations and properties for AIC, BIC and related criteria can be found in [157].

5.3.2.1 Information Criteria for DSS

Selecting a model using the GIC (5.37) approach requires the existence of a negative log-likelihood function which measures how well each candidate model fits the data. In the DSS framework, we estimate each model for a given κ by a penalised regression onto a smoothed data set $\bar{\mathbf{y}} = \mathbf{X}\bar{\boldsymbol{\beta}}$. Then model selection is equivalent to choosing the κ value that minimises a GIC type criterion. By construction, the unpenalised model in the regression path corresponding to $\kappa = 0$ will fit the data $\bar{\mathbf{y}}$ perfectly with no errors, resulting in an infinite log-likelihood. Therefore, a direct application of the GIC is not possible without some modification.

Consider the Kullback–Leibler (KL) [18] divergence from a probability distribution P to another probability distribution Q

$$D_{KL}(P\|Q) = -\mathbb{E}_P\left[\log\frac{q(x)}{p(x)}\right] = -\mathbb{E}_P\left[\log q(x)\right] + \mathbb{E}_P\left[\log p(x)\right], \tag{5.27}$$

where P is the generating, or "true" distribution of the data, and Q is some approximating distribution. The KL divergence measures the expected difference in negative-log likelihood between distribution P and Q, assuming the data was generated by P. Due to the similarity of the KL divergence to a negative log-likelihood, we propose the use of the KL divergence as a surrogate for the negative log-likelihood in the GIC (5.37). This yields a GIC-type criterion for selecting candidate models generated by the DSS procedure.

In order to use the KL divergence within the GIC framework, we require a reference model and an approximating model. Inspired by the original DSS procedure, we choose the reference model to be the Gaussian distribution with mean $\bar{\mathbf{y}}$ of (5.3) and variance $\bar{\sigma}^2 = \mathbb{E}\left[\sigma^2 \,|\, \mathbf{y}\right]$, i.e., the posterior mean of σ^2. Let $\boldsymbol{\beta}_\kappa$ denote the sparse coefficient vector found by solving (5.20) or an appropriate approximation to this optimisation problem such as (5.21); the generalised information score becomes

$$\mathrm{GIC}(\boldsymbol{\beta}_\kappa) = \min_{\sigma^2}\left\{D(\bar{\boldsymbol{\beta}}, \bar{\sigma}^2 \,\|\, \boldsymbol{\beta}_\kappa, \sigma^2) + \alpha(k_\kappa, \sigma^2)\right\}, \tag{5.28}$$

where

$$D(\bar{\boldsymbol{\beta}}, \bar{\sigma}^2 \,\|\, \boldsymbol{\beta}_\kappa, \sigma^2) = \frac{n}{2} \log \frac{\sigma^2}{\bar{\sigma}^2} - \frac{n}{2} + \frac{n^2 \bar{\sigma}^2}{2\sigma^2} + \frac{\|\mathbf{X}\boldsymbol{\beta}_\kappa - \mathbf{X}\bar{\boldsymbol{\beta}}\|_2^2}{2\sigma^2},$$

is the KL divergence between the reference model, $N(\mathbf{X}\bar{\boldsymbol{\beta}}, \bar{\sigma}^2)$ and the candidate sparse model $N(\mathbf{X}\boldsymbol{\beta}_\kappa, \sigma^2)$, and k_κ is the degrees of freedom of the sparse regression coefficient vector $\boldsymbol{\beta}_\kappa$, which is determined by the particular sparsification scheme that has been employed. Using this approach, the best sparsified regression coefficient vector $\boldsymbol{\beta}_\kappa$ is the one that minimises (5.28) among all candidate values of κ. The optimal solution to the group DSS procedure is achieved by balancing the trade-off between the number of predictors in the model and the closeness of the sparse model to the estimated model based on the posterior mean.

In our definition of the modified GIC (5.28) we treat the noise variance parameter σ^2 of our sparse model as a nuisance parameter and estimate it by minimisation. For a given sparse coefficient vector $\boldsymbol{\beta}_\kappa$, the value of σ^2 that minimises the GIC is given by

$$\hat{\sigma}^2 = \frac{\|\mathbf{X}\boldsymbol{\beta}_\kappa - \mathbf{X}\bar{\boldsymbol{\beta}}\|^2}{n - d} + \bar{\sigma}^2,$$

where $d = k_\kappa$ for the MML_u criterion (5.24) and $d = 0$ otherwise.

This information criteria-based approach has the advantage of allowing us to utilise the substantial body of work that has been undertaken on the problem of variable selection. Further, it provides an objective method for choosing an appropriate degree of posterior sparsification that, in comparison to the heuristic proposed in the original DSS paper, does not require any user-specified parameters. This approach can be used in both the original DSS procedure, as well as in our new group DSS extension to select an appropriate sparse model.

5.3.2.2 An Alternative Estimate for the Degrees of Freedom

The information criteria approach for selecting a sparse model requires an estimate of the degrees of freedom for all candidate models. In the standard DSS which is based on ℓ_1 penalisation, the degrees of freedom of each model can be approximated by the number of non-zero regression coefficients. Recall that, in the group DSS

procedure proposed in Section 5.3.1, we approximate the ideal ℓ_0 minimisation problem (5.20) by the more tractable group nonnegative garrotte estimator (5.21). An estimate for the degrees of freedom (df) of the group NNG was proposed by [37]

$$\widehat{\mathrm{df}}_{\mathrm{YL}}(\boldsymbol{\beta}_\kappa) = 2\sum_g \mathbb{1}(d_g > 0) + \sum_g d_g(s_g - 2). \tag{5.29}$$

This estimate of the degrees of freedom does not take into account within-group sparsity and will tend to overestimate the effective degrees of freedom for groups where some predictors are inactive. The group DSS procedure uses a smoothed version of the response variable $\mathbf{y}$, $\bar{\mathbf{y}}$, as the target for the subsequent group NNG sparsification step. Using $\bar{\boldsymbol{\beta}}$ in (5.21) takes into account the coefficient shrinkage induced by the continuous shrinkage prior distributions. Coefficients that are heavily shrunk contribute less than a full degree of freedom; therefore, the effective degrees of freedom for the model are less than suggested by (5.29) and should be adjusted in accordance with the degree of shrinkage. An overestimation of the degrees of freedom will result in an increased probability of erroneously rejecting large groups that contain only a few active coefficients.

As an illustration, consider a sparse group containing $s_g = 100$ predictors of which only a few are associated with the target. The estimate of the degrees of freedom given by (5.29) assumes that all 100 variables within the group are active. However, the target $\bar{\mathbf{y}}$ used in the information criteria is formed from the posterior mean $\bar{\boldsymbol{\beta}}$. After shrinkage, unassociated variables within the group will have posterior mean coefficient estimates that are close to zero and have little or no influence on the smoothed data $\bar{\mathbf{y}}$. This means that the information criteria penalty terms for models containing large groups where some predictors are inactive will be much larger than if the posterior shrinkage was taken into account.

We propose an alternative estimate of the degrees of freedom that takes into account the degree of coefficient shrinkage within all groups. The key idea is to exploit the form of the Bayesian linear regression hierarchy (5.8) which allows us to calculate the posterior expected degrees of freedom. Recall that the hierarchical

representation of the global-local shrinkage model is

$$\begin{aligned} \mathbf{y} \,|\, \mathbf{X}, \boldsymbol{\beta}, \sigma^2 &\sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I}_n), \\ \boldsymbol{\beta} \,|\, \sigma^2, \tau^2, \boldsymbol{\lambda}, \boldsymbol{\delta} &\sim N(\mathbf{0}, \sigma^2 \tau^2 \mathbf{D}_{\boldsymbol{\lambda}} \mathbf{D}_{\boldsymbol{\delta}_1} \cdots \mathbf{D}_{\boldsymbol{\delta}_K}). \end{aligned}$$

Conditional on τ, $\boldsymbol{\delta}$ and $\boldsymbol{\lambda}$, the above hierarchy reduces to a generalised ridge regression model. The posterior mean of $\boldsymbol{\beta}$ for this hierarchy is

$$\bar{\boldsymbol{\beta}}_{\tau^2,\boldsymbol{\lambda},\boldsymbol{\delta}} = \mathbb{E}\left[\boldsymbol{\beta} \,|\, \mathbf{y}, \tau^2, \boldsymbol{\lambda}, \boldsymbol{\delta}\right] = \left(\mathbf{X}^T\mathbf{X} + \boldsymbol{\Sigma}_{\tau^2,\boldsymbol{\lambda},\boldsymbol{\delta}}\right)^{-1} \mathbf{X}^T\mathbf{y} \tag{5.30}$$

where $\boldsymbol{\Sigma}_{\tau^2,\boldsymbol{\lambda},\boldsymbol{\delta}} = (\tau^2 \mathbf{D}_{\boldsymbol{\lambda}} \mathbf{D}_{\boldsymbol{\delta}_1} \cdots \mathbf{D}_{\boldsymbol{\delta}_K})^{-1}$ and the matrices $\mathbf{D}_{\boldsymbol{\lambda}}$ and $\mathbf{D}_{\boldsymbol{\delta}_k}$ are defined in Section 5.2.1. The effective degrees of freedom of a ridge regression model, conditional on the hyper-parameters $\tau^2, \boldsymbol{\lambda}, \boldsymbol{\delta}$, is [158]

$$\mathrm{df}(\bar{\boldsymbol{\beta}}_{\tau^2,\boldsymbol{\lambda},\boldsymbol{\delta}}) = \mathrm{tr}\left(\mathbf{X}(\mathbf{X}^T\mathbf{X} + \boldsymbol{\Sigma}_{\tau^2,\boldsymbol{\lambda},\boldsymbol{\delta}})^{-1}\mathbf{X}^T\right). \tag{5.31}$$

This estimate of the degrees of freedom depends on the particular values of the hyper-parameters τ^2, $\boldsymbol{\lambda}$ and $\boldsymbol{\delta}$; to remove this dependency we average (5.31) over the posterior distribution of τ^2, $\boldsymbol{\lambda}$ and $\boldsymbol{\delta}$ which yields the posterior expected degrees of freedom:

$$\mathrm{df}(\bar{\boldsymbol{\beta}}) = \mathbb{E}_{\tau^2,\boldsymbol{\lambda},\boldsymbol{\delta}}\left[\mathrm{tr}\left(\mathbf{X}(\mathbf{X}^T\mathbf{X} + \boldsymbol{\Sigma}_{\tau^2,\boldsymbol{\lambda},\boldsymbol{\delta}})^{-1}\mathbf{X}^T\right) \,|\, \mathbf{y}\right]. \tag{5.32}$$

This equation yields an estimate for the degrees of freedom of the full regression model which can be used to compute the information criterion score for any candidate model in the sparsification path produced by the group NNG. However, when the number of candidate models is large, this approach is computationally challenging. An approximation to (5.32) that is less computationally expensive can be found by decomposing the degrees of freedom of the full regression model into the sum of the degrees of freedom of each of the G groups. In the case that all groups are mutually orthogonal, this approximation is equivalent to the degrees of freedom of the full regression model (5.32). We therefore require an estimate of the degrees of freedom for each of the G groups that comprise the full model. As discussed in Section 5.3.1, we restrict our attention to models with a single level

of G non-overlapping groups. In this case, the degrees of freedom for a group g is

$$\mathrm{df}(\bar{\boldsymbol{\beta}}_g) = \mathbb{E}\left[\mathrm{tr}\left(\mathbf{X}_g\left(\mathbf{X}_g^T\mathbf{X}_g + \boldsymbol{\Sigma}_g\right)^{-1}\mathbf{X}_g^T\right) \mid \mathbf{y}\right], \tag{5.33}$$

where $\boldsymbol{\Sigma}_g = (\tau^2\mathbf{D}_{\boldsymbol{\lambda}_g}\mathbf{D}_{\boldsymbol{\delta}_g})^{-1}$. Using (5.33), our estimate for the degrees of freedom for a candidate model $\boldsymbol{\beta}_\kappa$ in the group NNG path is

$$\widehat{\mathrm{df}}_{\mathrm{PE}}(\boldsymbol{\beta}_\kappa) = \sum_{g=1}^{G} \mathbb{1}(d_g \neq 0)\cdot \mathrm{df}(\bar{\boldsymbol{\beta}}_g). \tag{5.34}$$

In practice, to calculate the expectation in (5.33), we use the samples from the posterior distribution and evaluate the equivalent expression

$$\mathrm{df}(\bar{\boldsymbol{\beta}}_g) = \mathrm{tr}\left(\mathbf{X}_g^T\mathbf{X}_g\,\mathbb{E}\left[\left(\mathbf{X}_g^T\mathbf{X}_g + \boldsymbol{\Sigma}_g\right)^{-1} \mid \mathbf{y}\right]\right), \tag{5.35}$$

which requires only one evaluation of $\mathbf{X}_g^T\mathbf{X}_g$, and involves less computation than (5.33).

5.4 Results of Simulated Examples

5.4.1 Example 1

We compared the performance of our Bayesian sparse group selection using continuous shrinkage priors against the Bayesian group spike-and-slab approach (BSGS) in [77] (see Section 2.4.1). The BSGS approach has been shown to outperform the standard group variable selection method, the sparse group Lasso [10], in simulation studies, and can be viewed as a potential "gold standard" in Bayesian grouped inference. Our simulations followed the experimental setup of [77]. In Example 1 there were $n = 50$ observations and $p = 60$ predictors. The predictors were divided into six groups, in which the first two groups contained 5 variables each, the second two groups contained 10 variables each and the last two groups contained 15 variables each. There were a total of 5 active predictors with coefficients $\beta_3 = 3.2$, $\beta_{11} = -2$, $\beta_{12} = 1$, $\beta_{31} = 1.5$, $\beta_{32} = -1.5$. The design matrix,

$\mathbf{X}$, was generated from the multivariate normal distribution with a mean of 0, a variance of 1.25, and a correlation of 0.2 between any two variables in the same group and 0 otherwise. The variance of the errors, σ^2, was varied between 4 to 2, resulting in a signal-to-noise ratio of 6.74 and 12.49, respectively.

We used the grouped Bayesian horseshoe prior (see Section 5.2.3) to obtain 10,000 posterior samples with the first 1,000 samples discarded as burn-in. The sampling procedure was done using the 'BayesReg' software toolbox [69] which implements the grouped global-local shrinkage priors detailed in this paper; this can be downloaded from MathWorks File Exchange (ID: 60823). The mean of the posterior samples was then used in the group non-negative garrote to obtain the solution path $\boldsymbol{\beta}_\kappa$. Four information criteria, BIC, AIC, AIC_c and MML_u (see Section 5.3.2), were computed for each model in the path using two different estimates of the degrees of freedom: (1) the conventional group NNG degrees of freedom given by (5.29), and (2) the posterior expected degrees of freedom given by (5.34), which we called BIC^*, AIC^*, AIC_c^* and MML_u^*. For the AIC_c criterion, if the estimated degrees of freedom is greater than the number of observations n, the denominator of the second term in the AICc formula (5.26) is negative and the corresponding candidate models are not considered for selection. The final sparse estimates were chosen by minimisation of the four information criteria. As a comparison, we also calculated the variation explained and excess error for each sparse model in the grouped NNG path (see Section 5.1.2). We selected the sparsest model for which the 90% credible interval of the variation explained and excess error statistics contained the expected value of the variation explained and excess error when κ is zero [14], respectively.

Since the datasets used for this example in [77] were not provided, we instead generated $10,000$ datasets using the procedure described above and recorded the number of times each method correctly identified groups as active or inactive. Due to the very long run times of the BSGS procedure, we did not include the BSGS method in these simulations; instead, we used the results reported in [77] for comparison and scale our results to match the number of simulations in [77].

The results for the simulations in Example 1, when $\sigma^2 = 4$, are shown in Table 5.1. In terms of the overall sum of correct identifications, the methods with estimated degrees of freedom $\widehat{\mathrm{df}}_{\mathrm{YL}}$ frequently failed to select active sparse groups (i.e., most predictors within the group are unassociated with the target) that were associated with the target. In contrast, our proposed degrees of freedom $\widehat{\mathrm{df}}_{\mathrm{PE}}$ showed a great improvement compared with $\widehat{\mathrm{df}}_{\mathrm{YL}}$ for BIC, AIC_c and MML_u. This was particularly apparent for group 5, which is an extremely sparse group with only two of the 15 predictors active. Our proposed method produced a much smaller estimate of the degrees of freedom and therefore resulted in higher probabilities of correctly selecting active sparse groups.

Although AIC with $\widehat{\mathrm{df}}_{\mathrm{PE}}$ (i.e., AIC^* in Table 5.1) correctly identified more active sparse groups than AIC with $\widehat{\mathrm{df}}_{\mathrm{YL}}$, it selected more inactive groups; as the AIC is known to overfit, particularly for small sample sizes n or when the number of predictors, p, is close to n, this increased rate of false positives is likely due to the AIC approximations being inaccurate in this setting. The AIC_c using $\widehat{\mathrm{df}}_{\mathrm{YL}}$ did not perform well at selecting the active sparse groups in this example as it tended to overestimate the degrees of freedom associated with these groups. However, AIC_c with our proposed estimate $\widehat{\mathrm{df}}_{\mathrm{PE}}$ appeared to more accurately estimate the effective degrees of freedom, and therefore tended to correctly select the active sparse groups more frequently. The BIC^*, AICc^* and MMLu^* had very competitive performance in comparison with the original BSGS procedure, while being substantially quicker to run.

TABLE 5.1: Percentage of times the six groups are selected in Example 1 for $\sigma^2 = 4$ by Bayesian sparse group selection model (BSGS), variation explained (VarExp), excess error (ExcErr), information criteria approaches with degrees of freedom $\widehat{\mathrm{df}}_{\mathrm{YL}}$ (BIC, AIC, AIC_c, MML_u), and information criteria approaches with posterior expected degrees of freedom $\widehat{\mathrm{df}}_{\mathrm{PE}}$ (BIC^*, AIC^*, AIC_c^*, MML_u^*). Active groups are marked in bold.

Group	Active No. of Variables /Group Size	BSGS	VarExp	ExcErr	BIC	AIC	AIC_c	MML_u	BIC^*	AIC^*	AIC_c^*	MML_u^*
1	1/5	100	100.0	100.0	99.7	100.0	100.0	100.0	100.0	100.0	100.0	100.0
2	0/5	7	1.8	2.5	0.4	2.8	0.1	0.2	8.4	18.6	8.5	6.1
3	2/10	100	97.6	97.3	66.5	96.9	65.8	77.1	99.3	99.8	99.6	99.1
4	0/10	1	1.4	2.3	0.1	1.7	0.0	0.0	5.5	14.8	5.2	3.5
5	2/15	97	89.7	90.3	39.9	87.3	7.1	41.7	95.8	98.5	97.3	94.8
6	0/15	2	1.1	2.0	0.1	0.9	0.0	0.0	3.9	12.0	3.4	2.1
	Overall sum of correct identifications	587	583.0	580.9	505.6	578.8	472.8	518.6	577.2	552.8	579.8	582.1

The results for $\sigma^2 = 2$, i.e. a larger signal-to-noise ratio, are shown in Table 5.2. In this case, the level of noise is low and the active groups are easier to correctly identify. From the table, the MML and AIC_c criteria using $\widehat{\mathrm{df}}_{\mathrm{PE}}$ showed better performance in selecting sparse groups (group 3 and 5) compared to the same criteria using $\widehat{\mathrm{df}}_{\mathrm{YL}}$. In terms of the overall sum of correct identifications, the MML_u^*, variation explained and excess error criteria were all virtually indistinguishable from the BSGS method, with AIC_c^* being only slightly worse. The BIC and AIC criteria with $\widehat{\mathrm{df}}_{\mathrm{YL}}$ performed better than the same criteria with $\widehat{\mathrm{df}}_{\mathrm{PE}}$; this appears to be due the inaccuracy of the asymptotic approximations used in these methods, meaning that the conservative estimate $\widehat{\mathrm{df}}_{\mathrm{YL}}$ of the degrees of freedom artificially assisted in preventing overfitting. This was supported by the fact that the small sample size versions of these criteria (i.e., MML_u^* and AIC_c^*) performed substantially better with our proposed estimate of the degrees of freedom.

TABLE 5.2: Percentage of times the six groups are selected in Example 1 for $\sigma^2 = 2$ by Bayesian sparse group selection model (BSGS), variation explained (VarExp), excess error (ExcErr), information criteria approaches with degrees of freedom $\widehat{\mathrm{df}}_{\mathrm{YL}}$ (BIC, AIC, AIC_c, MML_u), and information criteria approaches with posterior expected degrees of freedom $\widehat{\mathrm{df}}_{\mathrm{PE}}$ (BIC^*, AIC^*, AIC_c^*, MML_u^*). Active groups are marked in bold.

Group	Active No. of Variables /Group Size	BSGS	VarExp	ExcErr	BIC	AIC	AICc	MML_u	BIC^*	AIC^*	AIC_c^*	MML_u^*
1	1/5	100	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
2	0/5	2	1.0	1.8	0.7	2.6	0.0	0.0	6.8	16.8	6.0	3.2
3	2/10	100	100.0	100.0	98.0	100.0	79.6	95.6	100.0	100.0	100.0	100.0
4	0/10	1	0.8	1.7	0.1	1.7	0.0	0.0	4.6	13.7	3.7	1.8
5	2/15	100	99.7	99.8	94.1	99.7	20.6	82.0	99.9	100.0	100.0	99.8
6	0/15	0	0.7	1.6	0.1	0.9	0.0	0.0	3.3	11.5	2.0	0.8
	Overall sum of correct identifications	597	597.1	594.6	591.2	594.4	500.2	577.5	585.3	557.9	588.2	594.0

For this example, we ran further simulations for $\sigma^2 = 8$ and $\sigma^2 = 16$ (results not shown). When the level of noise was high, our proposed methods with $\widehat{\mathrm{df}}_{\mathrm{PE}}$ generally selected active sparse groups more frequently than the same criteria using $\widehat{\mathrm{df}}_{\mathrm{YL}}$, and the AIC_c^* and MML_u^* criteria produced a greater overall sum of correct identifications compared to the other methods tested, including the variation explained and excess error criteria. From all the criteria tested, the MML_u^* criterion appeared to have the best overall performance, in terms of speed and accuracy.

5.4.2 Example 2

In Example 2, there were $n = 200$ observations and $p = 100$ predictors divided into 10 groups of equal size. The number of non-zero coefficients in the first six groups was 10, 8, 6, 4, 2, and 1, respectively. The non-zero coefficients were randomly sampled from the set $\{-1, 1\}$. The noise was sampled from independent and identically distributed normal distributions with mean zero and variance 16. In this case, the signal-to-noise ratio was 3.42. We ran the example for 100 iterations using the data provided by [77] and recorded the number of times that each group was correctly identified. The results are presented in Table 5.3.

TABLE 5.3: Number of times the ten groups are selected in Example 2 by Bayesian sparse group selection model (BSGS), variation explained (VarExp), excess error (ExcErr), information criteria approaches with degrees of freedom $\widehat{\mathrm{df}}_{\mathrm{YL}}$ (BIC, AIC, AIC_c, MML_u), and information criteria approaches with posterior expected degrees of freedom $\widehat{\mathrm{df}}_{\mathrm{PE}}$ (BIC^*, AIC^*, AIC_c^*, MML_u^*). Active groups are marked in bold.

Group	Active No. of Variables /Group Size	BSGS	VarExp	ExcErr	BIC	AIC	AICc	MML_u	BIC^*	AIC^*	AIC_c^*	MML_u^*
1	10/10	100	100	100	100	100	100	100	100	100	100	100
2	8/10	100	100	100	96	100	100	100	100	100	100	100
3	6/10	100	96	99	88	99	97	96	99	100	100	100
4	4/10	99	90	96	72	96	93	93	97	100	99	99
5	2/10	94	53	70	26	71	64	56	79	88	87	83
6	1/10	79	19	46	3	45	24	21	57	73	66	62
7	0/10	10	1	3	0	3	1	1	3	14	7	5
8	0/10	12	2	4	0	4	2	2	6	12	8	6
9	0/10	11	0	2	0	1	0	0	3	18	10	7
10	0/10	14	0	0	0	0	0	0	4	15	6	4
	Overall sum of correct identifications	925	855	902	785	903	875	863	916	902	921	922

The BIC, AIC_c and MML_u criteria using the posterior expected degrees of freedom $\widehat{\mathrm{df}}_{\mathrm{PE}}$ show improved rates of selection of the sparse active groups, and are substantially better in terms of overall numbers of correct identifications, in comparison to the same criteria using the conventional degrees of freedom $\widehat{\mathrm{df}}_{\mathrm{YL}}$. The overall sum of correct identifications for BIC^*, AIC_c^* and particularly MML_u^*, were similar to the results obtained by the BSGS method. While BIC^*, AIC_c^* and MML_u^* produced a competitive overall sum of correct identifications, they also picked less inactive groups than the BSGS did. Both the excess error, and particularly the variance explained criteria appear inferior to BSGS and BIC^*, AICc* and MMLu* in this setting. These criteria appear overly conservative, and achieve a low rate of correct identification of the sparsest groups (5 and 6).

While the results of the proposed group DSS procedure were highly competitive compared to the slab-and-spike BSGS approach in terms of the correct identifications of the groups, the computational cost of our procedure was significantly lower. For example, it took approximately 30 hours to run the BSGS for a single simulation with the prior variance for active variables τ varied over $\{0.5, 1\}$; in

contrast, our group DSS method using the grouped horseshoe prior took approximately 5 seconds to finish a single simulation, and approximately 500 seconds to complete all 100 simulations.

5.4.3 Example 3

For Example 3, simulations and the set-up procedure followed the experimental setup of Example 4 in [77]. In Example 3, there were $n = 100$ observations and $p = 1200$ predictors. The predictors were divided into six groups of equal size, in which each group contained 200 variables. There were three active predictors with coefficients $\beta_{11} = 1$, $\beta_{12} = 2$, $\beta_{13} = 4$, so only the first group was active. The design matrix, $\mathbf{X}$, was generated using same procedure from Example 1 and the variance of the errors, σ^2, was set to 2.

We used the grouped Bayesian horseshoe prior to obtain 10,000 posterior samples with the first 1,000 samples discarded as burn-in. The sampling procedure was same as before and the mean of the posterior samples was obtained. Then we used in the group non-negative garrote without specifying group structure to obtain the solution path $\boldsymbol{\beta}_\kappa$ for all 1200 predictors in the model. Next, four information criteria, BIC, AIC, AIC_c and MML_u were computed for each model in the path using both estimates of the degrees of freedom: $\widehat{\text{df}}_{\text{YL}}$ and $\widehat{\text{df}}_{\text{PE}}$, which were denoted as BIC^*, AIC^*, AIC_c^* and MML_u^*. The final sparse estimates were chosen by minimisation of the four information criteria.

We ran this example for 10 iterations, recorded the total number of times that those three true predictors were correctly identified and also recorded the total number of predictors that were selected. The results from both BSGS and our proposed method are presented in Table 5.4. From the table, all criteria using the posterior expected degrees of freedom $\widehat{\text{df}}_{\text{PE}}$ performed best among all method in terms of selection of true predictors because they selected three true predictors for all 10 iterations, followed by BSGS method which selected true predictors 29 times, and then four criteria using conventional degrees of freedom $\widehat{\text{df}}_{\text{YL}}$ performed worse in

terms of the selection of true predictors. However, all information criteria tend to over-select the total number of predictors. The BSGS only selected 46 predictors in total, while MML_u^* selected 177 predictors and MML_u selected 414 predictors in total. Compared with VarExp and ExcErr which selected 30 and 27 times for true predictors and selected 262 and 282 for total number of variables respectively, MML_u^* still performed better.

Table 5.5 shows the results of group selection for Example 3. The MML_u^* is the best method because it selected active group (group 1) all the time and selected none inactive group. The AIC* selected true group 10 times, but also selected inactive groups three times. Both BIC* and AICc* selected the true group 8 times, and they selected no false groups. The information criteria using $\widehat{\text{df}}_{\text{YL}}$ did not perform well in terms of group selection since they tended to select empty models. Although VarExp and ExcErr selected the active group all the time, it selected a few inactive groups. In total, the MMLu* gave the largest number of overall sum of correct identification.

TABLE 5.4: Total umber of true predictors selected and the total number of predictors selected for 10 iterations in Example 3 by Bayesian sparse group selection model (BSGS), variation explained (VarExp), excess error (ExcErr), information criteria approaches with degrees of freedom $\widehat{\text{df}}_{\text{YL}}$ (BIC, AIC, AIC_c, MML_u), and information criteria approaches with posterior expected degrees of freedom $\widehat{\text{df}}_{\text{PE}}$ (BIC*, AIC*, AIC_c^*, MML_u^*).

Methods	BSGS	VarExp	ExcErr	BIC	AIC	AIC_c	MML_u	BIC	AIC*	AIC_c^*	MML_u^*
No. of true predictors selected	29	30	27	22	24	24	22	30	30	30	30
No. of total predictors selected	46	268	282	444	562	535	414	350	494	371	177

TABLE 5.5: Number of times the six groups are selected in Example 3 by variation explained (VarExp), excess error (ExcErr), information criteria approaches with degrees of freedom $\widehat{\mathrm{df}}_{\mathrm{YL}}$ (BIC, AIC, AIC_c, MML_u), and information criteria approaches with posterior expected degrees of freedom $\widehat{\mathrm{df}}_{\mathrm{PE}}$ (BIC^*, AIC^*, AIC_c^*, MML_u^*). Active groups are marked in bold.

Group	Active No. of Variables /Group Size	VarExp	ExcErr	BIC	AIC	AICc	MML_u	BIC^*	AIC^*	AIC_c^*	MML_u^*
1	3/200	10	10	0	1	0	0	8	10	8	10
2	0/200	2	3	0	0	0	0	0	1	0	0
3	0/200	3	2	0	0	0	0	0	0	0	0
4	0/200	2	2	0	0	0	0	0	1	0	0
5	0/200	3	3	0	0	0	0	0	1	0	0
6	0/200	2	3	0	0	0	0	0	0	0	0
	Overall sum of correct identifications	48	47	50	51	50	50	58	57	58	60

5.5 Further Extensions

We have shown the performance of our proposed method for simulated datasets. The proposed procedure using our estimate of the degrees of freedom performed well in selecting active groups, especially when there is a high level of sparsity within groups as shown in Example 1 and 2. For both examples, the level of sparsity at group level is moderate. Furthermore, our method demonstrated similar performance to the Bayesian grouped slab-and-spike (BSGS) method in terms of the overall rate of correct group identifications, while requiring substantially less computational resources.

For example 3, there are only three active predictors out of total 1200 predictors in the model, all of them are within the first group. To select the individual predictors, we specified group structures in the sampling procedure and treated all variables in the model as one group in the group NNG step. Our method using proposed estimate of degrees of freedom selected true predictors all the time, but it selected more inactive predictors compared to BSGS. Among four information criteria procedures, MML_u^* outperformed both VarExp and ExcErr. In this section, we will further extend the previous proposed method by varying different parameters to test the performance.

In Section 5.3.2, we have introduced our proposed method for calculating different information criteria to select the final model. In this section, we will explore the proposed method by introducing three variations of this approach: coefficient estimates, likelihood function and the penalty term, to compare their performance. Specifically, we propose to (1) use different estimates of the coefficients; (2) utilise the exact likelihood in the information criteria, and (3) use additional penalty terms in the likelihood. The detailed variations are explained as below:

Variation.1 Recall the GIC is of the form

$$\text{GIC}(\boldsymbol{\beta}, \sigma^2) = L(\boldsymbol{\beta}, \sigma^2) + \alpha(k, n, \sigma^2), \tag{5.36}$$

where $\boldsymbol{\beta}$ is the coefficient estimates, $L(\cdot)$ is the likelihood function and $\alpha(\cdot)$ is the penalty term.

For the estimates of model coefficients $\boldsymbol{\beta}$, instead of directly using $\widehat{\beta}_{NNG}$ in the information criteria, we can also use the conventional least squares estimate $\widehat{\beta}_{LS}$ in computing the likelihood function. For nonzero predictors in the NNG path, we computed the least squares estimate for those predictors. So it is the solution that minimises the sum of squared residuals and it has a smaller discrepancy between the data and an estimated model. Therefore it is a more accurate estimate compared to $\widehat{\beta}_{NNG}$.

Variation.2 The usual formulation of the generalised information criteria consists of a likelihood function that measures the goodness-of-fit of a model under consideration, and a model complexity penalty term that is a function of the degrees of freedom of the model. The conventional approach of computing information criteria score is to use the negative log-likelihood function. Therefore, the second variation we consider is substituting KL with traditional likelihood function.

Variation.3 Under the case of small-n-large-p scenario, the traditional criteria such as AIC, BIC tend to choose too many variables. Therefore, Chen and Chen [159] proposed an extended Bayes information criteria by introducing an extra

penalty term on top of the existing penalty. For model selection with large model spaces, the proposed extended version of GIC sacrifices a small loss in the positive selection, but it excessively controls the false discovery rate.

We apply this extended version to our GIC in the model selection part. The penalty term is of the form $\ln \binom{p}{k}$, where p is the total number of predictors and k is the total number of nonzero estimates from NNG path. Since we are adding an extra penalty term, the GIC becomes:

$$\text{GIC}(\boldsymbol{\beta}, \sigma^2) = L(\boldsymbol{\beta}, \sigma^2) + \alpha(k, n, \sigma^2) + \ln \binom{p}{k}. \tag{5.37}$$

The plot of the penalty term $\ln \binom{p}{k}$ is shown in Figure 5.2. From the plot, we see that it is a increasing function towards the middle of p values and then a decreasing function which implies that it penalises models heavier when k is half of p, and it almost puts no penalty when k is close to 0 or close to p. This captures the total number of different models; with more variety of the model parameters, the bigger the penalty term is. For example, when $p = 100$ and $k = 0$ or 1, there is only one unique model; when $k = 50$, the curve achieves maximum which means it achieves maximum number of different models. Furthermore, the maximum number of models gets bigger as p increases.

Therefore, the 8 methods we are going to test the performance are as follows:

1. (NNG, KL, IC)
2. (NNG, L, IC)
3. (LS, KL, IC)
4. (LS, L, IC)
5. (NNG, KL, EIC)
6. (NNG, L, EIC)

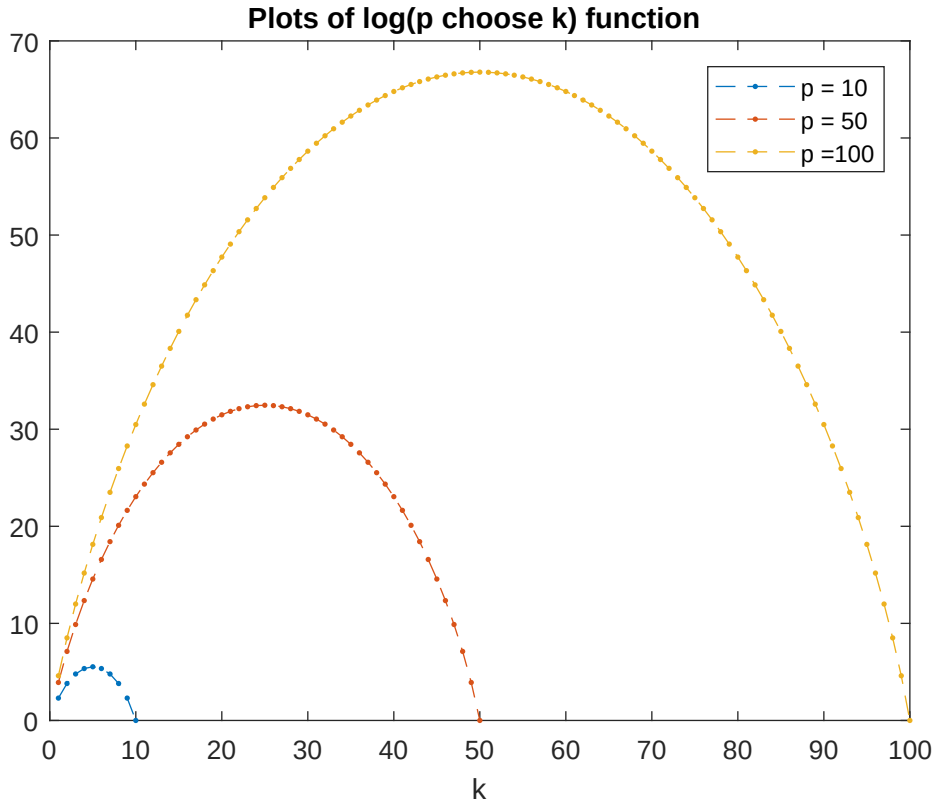

FIGURE 5.2: Plot of function $\ln \binom{p}{k}$ when $p = 10, 50, 100$.

7. (LS, KL, EIC)

8. (LS, L, EIC),

where NNG represents $\widehat{\beta}_{NNG}$ introduced in **Variation.1**, KL is the xxx introduced in **Variation.2** and EIC is the extended information criteria discussed in **Variation.3**. For example, (LS, KL, EIC) represents the approach that after we apply group DSS to posterior samples to reach a NNG solution path, we find the nonzero predictors in the NNG path, fit the least squares estimate to them and obtain LS estimates. The LS estimates then are substituted into the proposed GIC which consists of a KL term plus a penalty term $\ln \binom{p}{k}$ on top of existing penalty. Note that the first method (NNG, KL, IC) is simply our proposed method that we applied to all examples in Section 5.4.

5.5.1 Example 1

We rerun all simulated examples for the above 8 methods. The set up of Example 1 can be found in Section 5.4.1. Table 5.6 and Table 5.7 shows the results of all 8 methods when σ^2 is 4. From the tables, we see that (NNG, KL, IC), (LS, KL, IC), (LS, L, IC), (NNG, KL, EIC) and (LS, KL, EIC) performed relatively well compared to BSGS method, with (LS, L, IC) performed slightly worse than the rest four methods. This shows that KL approach outperformed likelihood approach.

When $\sigma^2 = 2$, results can be found in Table 5.8 and Table 5.9. The results suggest that (NNG, KL, IC), (LS, KL, IC), (LS, L, IC), (NNG, KL, EIC), (LS, KL, EIC), (LS, L, EIC) worked well in this case, with (LS, L, IC) and (LS, L, EIC) being slightly worse than the others.

TABLE 5.6: Number of times the six groups are selected in Example 1 for $\sigma^2 = 4$ for (NNG, KL, IC), (NNG, L, IC), (LS, KL, IC) and (LS, L, IC) by Bayesian sparse group selection model (BSGS), variation explained (VarExp), excess error (ExcErr), information criteria approaches with degrees of freedom $\widehat{\mathrm{df}}_{\mathrm{YL}}$ (BIC, AIC, AIC_c, MML_u), and information criteria approaches with posterior expected degrees of freedom $\widehat{\mathrm{df}}_{\mathrm{PE}}$ (BIC^*, AIC^*, AIC_c^*, MML_u^*). Active groups are marked in bold.

Method	Group	Active No. of Variables /Group Size	BSGS	VarExp	ExcErr	BIC	AIC	AICc	MML_u	BIC^*	AIC^*	AIC_c^*	MML_u^*
(NNG, KL, IC)	**1**	1/5	100	100	100	97	100	100	100	100	100	100	100
	2	0/5	7	2	2	0	1	0	0	2	8	4	2
	3	2/10	100	99	99	16	85	42	58	98	100	100	99
	4	0/10	1	0	0	0	0	0	0	1	5	1	1
	5	2/15	97	90	92	2	66	1	24	87	98	93	87
	6	0/15	2	1	2	0	0	0	0	1	3	0	0
		Overall sum of correct identifications	587	586	587	415	550	443	482	581	582	588	583
(NNG, L, IC)	**1**	1/5	100	100	100	98	100	100	100	100	100	100	100
	2	0/5	7	2	2	0	4	0	0	33	81	31	25
	3	2/10	100	99	99	30	99	48	65	100	100	100	100
	4	0/10	1	0	0	0	3	0	1	29	78	27	16
	5	2/15	97	90	92	14	89	4	45	100	100	100	100
	6	0/15	2	1	2	0	2	0	1	22	72	18	15
		Overall sum of correct identifications	587	586	587	442	579	452	508	516	369	524	544
(LS, KL, IC)	**1**	1/5	100	100	100	100	100	100	100	100	100	100	100
	2	0/5	7	2	2	0	1	0	0	1	2	2	2
	3	2/10	100	99	99	63	80	75	74	93	99	98	97
	4	0/10	1	0	0	0	0	0	0	0	0	0	0
	5	2/15	97	90	92	17	34	18	18	81	93	89	83
	6	0/15	2	1	2	0	0	0	0	0	0	0	0
		Overall sum of correct identifications	587	586	587	480	513	493	492	573	590	585	578
(LS, L, IC)	**1**	1/5	100	100	100	100	100	100	100	100	100	100	100
	2	0/5	7	2	2	1	5	1	2	16	38	18	11
	3	2/10	100	99	99	74	86	76	76	100	100	100	100
	4	0/10	1	0	0	0	2	0	0	17	37	18	10
	5	2/15	97	90	92	22	85	18	21	100	100	100	100
	6	0/15	2	1	2	0	6	0	0	10	30	8	3
		Overall sum of correct identifications	587	586	587	495	558	493	495	557	495	556	576

TABLE 5.7: Number of times the six groups are selected in Example 1 for $\sigma^2 = 4$ for (NNG, KL, EIC), (NNG, L, EIC), (LS, KL, EIC) and (LS, L, EIC) by Bayesian sparse group selection model (BSGS), variation explained (VarExp), excess error (ExcErr), information criteria approaches with degrees of freedom $\widehat{\mathrm{df}}_{\mathrm{YL}}$ (BIC, AIC, AIC_c, MML_u), and information criteria approaches with posterior expected degrees of freedom $\widehat{\mathrm{df}}_{\mathrm{PE}}$ (BIC^*, AIC^*, AIC_c^*, MML_u^*). Active groups are marked in bold.

Method	Group	Active No. of Variables /Group Size	BSGS	VarExp	ExcErr	BIC	AIC	AICc	MML_u	BIC^*	AIC^*	AIC_c^*	MML_u^*
(NNG, KL, EIC)	**1**	1/5	100	100	100	92	100	100	99	100	100	100	100
	2	0/5	7	2	2	0	1	0	4	3	8	4	4
	3	2/10	100	99	99	13	85	42	47	98	100	100	99
	4	0/10	1	0	0	0	0	0	0	1	5	1	3
	5	2/15	97	90	92	2	66	1	18	87	98	93	87
	6	0/15	2	1	2	0	0	0	4	1	3	0	2
		Overall sum of correct identifications	587	586	587	407	550	443	456	580	582	588	577
(NNG, L, EIC)	**1**	1/5	100	100	100	93	100	100	100	100	100	100	100
	2	0/5	7	2	2	0	4	0	0	81	81	31	74
	3	2/10	100	99	99	22	99	48	58	100	100	100	100
	4	0/10	1	0	0	0	3	0	1	79	78	27	74
	5	2/15	97	90	92	11	89	4	42	100	100	100	100
	6	0/15	2	1	2	0	2	0	1	78	72	18	72
		Overall sum of correct identifications	587	586	587	426	579	452	498	362	369	524	380
(LS, KL, EIC)	**1**	1/5	100	100	100	100	100	100	100	100	100	100	100
	2	0/5	7	2	2	0	1	0	0	1	2	2	1
	3	2/10	100	99	99	59	80	75	74	91	99	98	94
	4	0/10	1	0	0	0	0	0	0	0	0	0	0
	5	2/15	97	90	92	17	34	18	18	79	93	89	81
	6	0/15	2	1	2	0	0	0	0	0	0	0	0
		Overall sum of correct identifications	587	586	587	476	513	493	492	569	590	585	574
(LS, L, EIC)	**1**	1/5	100	100	100	100	100	100	100	100	100	100	100
	2	0/5	7	2	2	1	5	1	1	26	38	18	21
	3	2/10	100	99	99	72	86	76	76	100	100	100	100
	4	0/10	1	0	0	0	2	0	0	23	37	18	12
	5	2/15	97	90	92	22	85	18	20	100	100	100	100
	6	0/15	2	1	2	0	6	0	0	12	30	8	9
		Overall sum of correct identifications	587	586	587	493	558	493	495	539	495	556	558

TABLE 5.8: Number of times the six groups are selected in Example 1 for $\sigma^2 = 2$ for (NNG, KL, IC), (NNG, L, IC), (LS, KL, IC) and (LS, L, IC) by Bayesian sparse group selection model (BSGS), variation explained (VarExp), excess error (ExcErr), information criteria approaches with degrees of freedom $\widehat{\mathrm{df}}_{\mathrm{YL}}$ (BIC, AIC, AIC_c, MML_u), and information criteria approaches with posterior expected degrees of freedom $\widehat{\mathrm{df}}_{\mathrm{PE}}$ (BIC^*, AIC^*, AIC_c^*, MML_u^*). Active groups are marked in bold.

Method	Group	Active No. of Variables /Group Size	BSGS	VarExp	ExcErr	BIC	AIC	AICc	MML_u	BIC^*	AIC^*	AIC_c^*	MML_u^*
(NNG, KL, IC)	**1**	1/5	100	100	100	99	100	100	100	100	100	100	100
	2	0/5	2	1	1	0	1	0	0	2	7	2	1
	3	2/10	100	100	100	53	100	67	85	100	100	100	100
	4	0/10	1	0	0	0	0	0	0	0	4	0	0
	5	2/15	100	100	100	38	98	3	56	100	100	100	100
	6	0/15	0	1	1	0	0	0	0	1	3	0	0
		Overall sum of correct identifications	597	598	598	490	597	470	541	597	586	598	599
(NNG, L, IC)	**1**	1/5	100	100	100	99	100	100	100	100	100	100	100
	2	0/5	2	1	1	1	4	0	0	32	82	28	15
	3	2/10	100	100	100	83	100	68	93	100	100	100	100
	4	0/10	1	0	0	0	3	0	0	26	79	18	8
	5	2/15	100	100	100	78	100	3	74	100	100	100	100
	6	0/15	0	1	1	0	2	0	0	20	77	10	6
		Overall sum of correct identifications	597	598	598	559	591	471	567	522	362	544	571
(LS, KL, IC)	**1**	1/5	100	100	100	100	100	100	100	100	100	100	100
	2	0/5	2	1	1	0	0	0	0	1	1	0	0
	3	2/10	100	100	100	90	93	90	90	100	100	100	100
	4	0/10	1	0	0	0	0	0	0	0	0	0	0
	5	2/15	100	100	100	13	89	10	11	99	100	100	99
	6	0/15	0	1	1	0	0	0	0	0	0	0	0
		Overall sum of correct identifications	597	598	598	503	582	500	501	598	599	600	599
(LS, L, IC)	**1**	1/5	100	100	100	100	100	100	100	100	100	100	100
	2	0/5	2	1	1	0	4	0	0	18	39	17	5
	3	2/10	100	100	100	90	99	90	90	100	100	100	100
	4	0/10	1	0	0	0	5	0	0	20	44	19	5
	5	2/15	100	100	100	49	100	12	16	100	100	100	100
	6	0/15	0	1	1	1	4	0	0	9	28	8	1
		Overall sum of correct identifications	597	598	598	538	586	502	506	553	489	556	589

TABLE 5.9: Number of times the six groups are selected in Example 1 for $\sigma^2 = 2$ for (NNG, KL, EIC), (NNG, L, EIC), (LS, KL, EIC) and (LS, L, EIC) by Bayesian sparse group selection model (BSGS), variation explained (VarExp), excess error (ExcErr), information criteria approaches with degrees of freedom $\widehat{\mathrm{df}}_{\mathrm{YL}}$ (BIC, AIC, AIC_c, MML_u), and information criteria approaches with posterior expected degrees of freedom $\widehat{\mathrm{df}}_{\mathrm{PE}}$ (BIC^*, AIC^*, AIC_c^*, MML_u^*). Active groups are marked in bold.

Method	Group	Active No. of Variables /Group Size	BSGS	VarExp	ExcErr	BIC	AIC	AICc	MML_u	BIC^*	AIC^*	AIC_c^*	MML_u^*
(NNG, KL, EIC)	**1**	1/5	100	100	100	98	100	100	100	100	100	100	100
	2	0/5	2	1	1	0	1	0	0	3	7	2	2
	3	2/10	100	100	100	42	100	67	73	100	100	100	100
	4	0/10	1	0	0	0	0	0	0	0	4	0	0
	5	2/15	100	100	100	32	98	3	51	100	100	100	100
	6	0/15	0	1	1	0	0	0	0	1	3	0	0
		Overall sum of correct identifications	597	598	598	472	597	470	524	596	586	598	598
(NNG, L, EIC)	**1**	1/5	100	100	100	99	100	100	100	100	100	100	100
	2	0/5	2	1	1	1	4	0	0	78	82	28	32
	3	2/10	100	100	100	80	100	68	81	100	100	100	100
	4	0/10	1	0	0	0	3	0	0	76	79	18	24
	5	2/15	100	100	100	75	100	3	65	100	100	100	100
	6	0/15	0	1	1	0	2	0	0	76	77	10	22
		Overall sum of correct identifications	597	598	598	553	591	471	546	370	362	544	522
(LS, KL, EIC)	**1**	1/5	100	100	100	100	100	100	100	100	100	100	100
	2	0/5	2	1	1	0	0	0	0	1	1	0	0
	3	2/10	100	100	100	89	93	90	90	100	100	100	100
	4	0/10	1	0	0	0	0	0	0	0	0	0	0
	5	2/15	100	100	100	13	89	10	11	99	100	100	99
	6	0/15	0	1	1	0	0	0	0	0	0	0	0
		Overall sum of correct identifications	597	598	598	502	582	500	501	598	599	600	599
(LS, L, EIC)	**1**	1/5	100	100	100	100	100	100	100	100	100	100	100
	2	0/5	2	1	1	0	4	0	0	27	39	17	13
	3	2/10	100	100	100	90	99	90	90	100	100	100	100
	4	0/10	1	0	0	0	5	0	0	26	44	19	9
	5	2/15	100	100	100	48	100	12	14	100	100	100	100
	6	0/15	0	1	1	1	4	0	0	13	28	8	1
		Overall sum of correct identifications	597	598	598	537	586	502	504	534	489	556	577

5.5.2 Example 2

The set up of Example 2 can be found in Section 5.4.2. The results of Example 2 are shown in Table 5.10, Table 5.11 and Table 5.12. Method (NNG, KL, IC), (LS, KL, IC) and (LS, L, IC) are competitive to BSGS method. Method 3 identified less sparse active groups, but selected less false groups. Method 4 selected more active groups even when the groups are sparse, but it also selected more inactive groups.

TABLE 5.10: Number of times the ten groups are selected in Example 2 for (NNG, KL, IC), (NNG, L, IC) and (LS, KL, IC) by Bayesian sparse group selection model (BSGS), variation explained (VarExp), excess error (ExcErr), information criteria approaches with degrees of freedom $\widehat{\mathrm{df}}_{\mathrm{YL}}$ (BIC, AIC, AIC_c, MML_u), and information criteria approaches with posterior expected degrees of freedom $\widehat{\mathrm{df}}_{\mathrm{PE}}$ (BIC^*, AIC^*, AIC_c^*, MML_u^*). Active groups are marked in bold.

Method	Group	Active No. of Variables /Group Size	BSGS	VarExp	ExcErr	BIC	AIC	AICc	MML_u	BIC^*	AIC^*	AIC_c^*	MML_u^*
(NNG, KL, IC)	**1**	10/10	100	100	100	80	100	100	100	92	100	100	100
	2	8/10	100	100	100	71	99	99	99	88	100	100	100
	3	6/10	100	96	99	51	96	92	95	83	100	100	100
	4	4/10	99	89	96	22	88	75	83	75	99	98	97
	5	2/10	94	53	71	0	48	28	36	54	83	80	78
	6	1/10	79	20	47	0	15	3	7	31	61	57	58
	7	0/10	10	1	3	0	1	0	0	2	5	4	4
	8	0/10	12	2	4	0	2	0	0	2	6	4	5
	9	0/10	11	0	2	0	0	0	0	0	6	3	3
	10	0/10	14	0	0	0	0	0	0	0	4	3	4
		Overall sum of correct identifications	925	855	904	624	843	797	820	819	922	921	917
(NNG, L, IC)	**1**	10/10	100	100	100	92	100	100	100	100	100	100	100
	2	8/10	100	100	100	85	100	100	100	100	100	100	100
	3	6/10	100	96	99	73	100	98	96	100	100	100	100
	4	4/10	99	90	96	50	98	93	92	100	100	100	100
	5	2/10	94	53	71	11	82	63	63	89	100	99	96
	6	1/10	79	21	48	0	57	25	31	71	99	94	88
	7	0/10	10	1	3	0	5	1	1	13	82	54	44
	8	0/10	12	2	3	0	5	1	2	12	81	53	36
	9	0/10	11	0	2	0	3	0	0	17	84	58	40
	10	0/10	14	0	0	0	3	0	0	16	82	57	40
		Overall sum of correct identifications	925	857	906	711	921	877	879	902	670	771	824
(LS, KL, IC)	**1**	10/10	100	100	100	93	100	100	100	96	100	100	100
	2	8/10	100	100	100	82	99	99	97	90	100	100	100
	3	6/10	100	96	99	48	86	79	76	78	100	99	100
	4	4/10	99	90	95	26	74	61	53	58	95	93	92
	5	2/10	94	53	71	3	30	18	16	25	77	64	65
	6	1/10	79	19	48	1	7	5	4	4	53	43	43
	7	0/10	10	1	3	0	1	1	1	0	3	2	2
	8	0/10	12	2	4	0	0	0	0	0	4	1	1
	9	0/10	11	0	2	0	0	0	0	0	1	1	1
	10	0/10	14	0	0	0	0	0	0	0	0	0	0
		Overall sum of correct identifications	925	855	904	653	795	761	745	751	917	895	896

TABLE 5.11: Number of times the ten groups are selected in Example 2 for (LS, L, IC), (NNG, KL, EIC) and (NNG, L, EIC) by Bayesian sparse group selection model (BSGS), variation explained (VarExp), excess error (ExcErr), information criteria approaches with degrees of freedom $\widehat{\mathrm{df}}_{\mathrm{YL}}$ (BIC, AIC, AIC_c, MML_u), and information criteria approaches with posterior expected degrees of freedom $\widehat{\mathrm{df}}_{\mathrm{PE}}$ (BIC^*, AIC^*, AIC_c^*, MML_u^*). Active groups are marked in bold.

Method	Group	Active No. of Variables /Group Size	BSGS	VarExp	ExcErr	BIC	AIC	AICc	MML_u	BIC^*	AIC^*	AIC_c^*	MML_u^*
(LS, L, IC)	**1**	10/10	100	100	100	100	100	100	100	100	100	100	100
	2	8/10	100	100	100	93	100	100	100	100	100	100	100
	3	6/10	100	96	99	66	99	93	88	100	100	100	100
	4	4/10	99	89	96	42	87	82	70	99	100	100	100
	5	2/10	94	53	71	15	54	32	25	85	98	97	95
	6	1/10	79	20	47	2	22	9	7	57	98	91	80
	7	0/10	10	1	3	0	3	2	1	5	48	29	16
	8	0/10	12	2	4	0	6	1	0	5	47	24	19
	9	0/10	11	0	2	0	3	1	0	3	53	29	17
	10	0/10	14	0	0	0	0	0	0	1	52	29	16
		Overall sum of correct identifications	925	855	904	718	850	812	789	927	796	877	907
(NNG, KL, EIC)	**1**	10/10	100	100	100	68	100	100	100	89	100	100	100
	2	8/10	100	100	100	58	99	99	98	86	100	100	100
	3	6/10	100	96	99	37	96	92	94	81	100	100	100
	4	4/10	99	90	96	17	88	76	82	72	99	98	97
	5	2/10	94	53	71	0	49	28	35	54	83	80	87
	6	1/10	79	21	47	0	15	2	7	30	61	57	73
	7	0/10	10	1	3	0	1	0	0	2	5	4	45
	8	0/10	12	2	4	0	2	0	0	2	6	4	44
	9	0/10	11	0	2	0	0	0	0	0	5	3	43
	10	0/10	14	0	0	0	0	0	0	0	4	3	44
		Overall sum of correct identifications	925	857	904	580	844	797	816	808	923	921	781
(NNG, L, EIC)	**1**	10/10	100	100	100	84	100	100	100	100	100	100	100
	2	8/10	100	100	100	79	100	100	100	100	100	100	100
	3	6/10	100	96	99	68	100	98	96	100	100	100	100
	4	4/10	99	91	96	44	98	93	92	100	100	100	100
	5	2/10	94	53	72	10	82	63	64	92	100	99	100
	6	1/10	79	21	47	0	56	25	31	81	99	94	100
	7	0/10	10	1	3	0	5	1	1	41	82	55	100
	8	0/10	12	2	4	0	5	1	2	37	80	52	100
	9	0/10	11	0	2	0	2	0	0	42	84	58	100
	10	0/10	14	0	0	0	1	0	0	42	83	56	100
		Overall sum of correct identifications	925	858	905	685	923	877	880	811	670	772	600

TABLE 5.12: Number of times the ten groups are selected in Example 2 for (LS, KL, EIC) and (LS, L, EIC) by Bayesian sparse group selection model (BSGS), variation explained (VarExp), excess error (ExcErr), information criteria approaches with degrees of freedom $\widehat{\mathrm{df}}_{\mathrm{YL}}$ (BIC, AIC, AIC_c, MML_u), and information criteria approaches with posterior expected degrees of freedom $\widehat{\mathrm{df}}_{\mathrm{PE}}$ (BIC^*, AIC^*, AIC_c^*, MML_u^*). Active groups are marked in bold.

Method	Group	Active No. of Variables /Group Size	BSGS	VarExp	ExcErr	BIC	AIC	AICc	MML_u	BIC^*	AIC^*	AIC_c^*	MML_u^*
(LS, KL, EIC)	**1**	10/10	100	100	100	92	100	100	100	90	100	100	100
	2	8/10	100	100	100	81	99	99	97	85	100	100	100
	3	6/10	100	96	99	46	86	79	74	73	100	99	99
	4	4/10	99	90	95	23	74	61	50	54	95	93	92
	5	2/10	94	53	71	2	30	18	16	22	77	64	70
	6	1/10	79	19	48	1	7	5	4	2	53	43	48
	7	0/10	10	1	3	0	1	1	1	0	3	2	19
	8	0/10	12	2	4	0	0	0	0	0	4	1	18
	9	0/10	11	0	2	0	0	0	0	0	1	1	18
	10	0/10	14	0	0	0	0	0	0	0	0	0	17
		Overall sum of correct identifications	925	855	904	645	795	761	740	726	917	895	837
(LS, L, EIC)	**1**	10/10	100	100	100	100	100	100	100	100	100	100	100
	2	8/10	100	100	100	92	100	100	99	100	100	100	100
	3	6/10	100	96	99	63	99	94	86	100	100	100	100
	4	4/10	99	90	96	39	87	82	69	99	100	100	100
	5	2/10	94	53	71	13	56	33	26	85	98	97	100
	6	1/10	79	21	48	2	22	8	5	61	97	90	99
	7	0/10	10	1	3	0	1	1	1	4	47	28	98
	8	0/10	12	2	3	0	6	1	0	8	47	24	99
	9	0/10	11	0	2	0	3	1	0	5	52	28	98
	10	0/10	14	0	0	0	0	0	0	4	52	29	98
		Overall sum of correct identifications	925	857	906	709	854	814	784	924	797	878	606

5.5.3 Example 3

The set up of Example 3 can be found in Section 5.4.3. We performed both individual selection and group selection for this example.

The number of true predictors selected and the total number of predictors selected for all 8 methods are shown in Table 5.13. In this example, the sample size n is relatively small compared to the number of features p which is large. Four methods using IC selected true predictors all the time, but also selected more inactive predictors. Four methods using EIC selected much less inactive predictors compared to BSGS method, but they also selected true predictors 27 times in total.

From Table 5.14 and Table 5.15, it can be seen that all methods with our proposed degrees of freedom outperformed the methods with $\widehat{\mathrm{df}}_{\mathrm{YL}}$, as well as outperforming VarExp and ExcErr in terms of the overall sum of correct identification. In particular MMLu* selected the active sparse group all the time, and select none inactive group for all methods except (NNG, KL, EIC) and (NNG, L, EIC).

TABLE 5.13: Number of true predictors selected and the total number of predictors selected for 10 iterations in Example 3 for all 8 methods by Bayesian sparse group selection model (BSGS), variation explained (VarExp), excess error (ExcErr), information criteria approaches with degrees of freedom $\widehat{\mathrm{df}}_{\mathrm{YL}}$ (BIC, AIC, AIC_c, MML_u), and information criteria approaches with posterior expected degrees of freedom $\widehat{\mathrm{df}}_{\mathrm{PE}}$ (BIC^*, AIC^*, AIC_c^*, MML_u^*). Active groups are marked in bold.

Method	Title	BSGS	VarExp	ExcErr	BIC	AIC	AIC_c	MML_u	BIC	AIC^*	AIC_c^*	MML_u^*
(NNG, KL, IC)	No. of true predictors selected	29	30	27	22	24	24	22	30	30	30	30
	No. of predictors selected	46	268	282	444	562	535	414	350	494	371	177
(NNG, L, IC)	No. of true predictors selected	29	30	27	24	24	24	24	30	30	30	30
	No. of predictors selected	46	268	282	677	742	695	623	735	939	752	556
(LS, KL, IC)	No. of true predictors selected	29	30	27	22	24	24	22	30	30	30	30
	No. of predictors selected	46	268	282	440	538	510	415	309	432	343	178
(LS, L, IC)	No. of true predictors selected	29	30	27	24	24	24	24	30	30	30	30
	No. of predictors selected	46	268	282	672	690	623	621	601	705	612	449
(NNG, KL, EIC)	No. of true predictors selected	29	30	27	21	24	24	21	27	30	30	27
	No. of predictors selected	46	268	282	23	562	535	22	29	494	371	28
(NNG, L, EIC)	No. of true predictors selected	29	30	27	22	24	24	22	28	30	30	27
	No. of predictors selected	46	268	282	217	742	695	214	223	939	752	29
(LS, KL, EIC)	No. of true predictors selected	29	30	27	21	24	24	21	27	30	30	27
	No. of predictors selected	46	268	282	22	538	510	22	29	432	343	29
(LS, L, EIC)	No. of true predictors selected	29	30	27	22	24	24	22	28	30	30	27
	No. of predictors selected	46	268	282	211	690	623	211	216	705	612	29

TABLE 5.14: Number of times the six groups are selected in Example 3 for (NNG, KL, IC), (NNG, L, IC), (LS, KL, IC) and (LS, L, IC) by variation explained (VarExp), excess error (ExcErr), information criteria approaches with degrees of freedom $\widehat{\mathrm{df}}_{\mathrm{YL}}$ (BIC, AIC, AIC_c, MML_u), and information criteria approaches with posterior expected degrees of freedom $\widehat{\mathrm{df}}_{\mathrm{PE}}$ (BIC^*, AIC^*, AIC_c^*, MML_u^*). Active groups are marked in bold.

Method	Group	Active No. of Variables /Group Size	VarExp	ExcErr	BIC	AIC	AICc	MML_u	BIC^*	AIC^*	AIC_c^*	MML_u^*
(NNG, KL, IC)	**1**	3/200	10	10	0	0	0	0	8	10	8	10
	2	0/200	2	3	0	0	0	0	0	0	0	0
	3	0/200	3	2	0	0	0	0	0	0	0	0
	4	0/200	2	2	0	0	0	0	0	0	0	0
	5	0/200	3	3	0	0	0	0	0	0	0	0
	6	0/200	2	3	0	0	0	0	0	0	0	0
		Overall sum of correct identifications	48	47	50	50	50	50	58	60	58	60
(NNG, L, IC)	**1**	3/200	10	10	0	0	0	0	8	10	8	10
	2	0/200	2	3	0	0	0	0	0	4	0	0
	3	0/200	3	2	0	0	0	0	0	3	0	0
	4	0/200	2	2	0	0	0	0	0	3	0	0
	5	0/200	3	3	0	0	0	0	0	3	0	0
	6	0/200	2	3	0	0	0	0	0	3	0	0
		Overall sum of correct identifications	48	47	50	50	50	50	58	44	58	60
(LS, KL, IC)	**1**	3/200	10	10	0	3	0	0	10	10	8	10
	2	0/200	2	3	0	0	0	0	0	0	0	0
	3	0/200	3	2	0	0	0	0	0	0	0	0
	4	0/200	2	2	0	0	0	0	0	0	0	0
	5	0/200	3	3	0	0	0	0	0	0	0	0
	6	0/200	2	3	0	0	0	0	0	0	0	0
		Overall sum of correct identifications	48	47	50	53	50	50	60	60	58	60
(LS, L, IC)	**1**	3/200	10	10	2	6	0	0	10	10	8	10
	2	0/200	2	3	0	1	0	0	1	2	0	0
	3	0/200	3	2	0	0	0	0	0	1	0	0
	4	0/200	2	2	0	0	0	0	0	1	0	0
	5	0/200	3	3	0	1	0	0	1	1	0	0
	6	0/200	2	3	0	1	0	0	0	2	0	0
		Overall sum of correct identifications	48	47	52	53	50	50	58	53	58	60

TABLE 5.15: Number of times the six groups are selected in Example 3 for (NNG, KL, EIC), (NNG, L, EIC), (LS, KL, EIC) and (LS, L, EIC) by variation explained (VarExp), excess error (ExcErr), information criteria approaches with degrees of freedom $\widehat{\mathrm{df}}_{\mathrm{YL}}$ (BIC, AIC, AIC_c, MML_u), and information criteria approaches with posterior expected degrees of freedom $\widehat{\mathrm{df}}_{\mathrm{PE}}$ (BIC^*, AIC^*, AIC_c^*, MML_u^*). Active groups are marked in bold.

Method	Group	Active No. of Variables /Group Size	VarExp	ExcErr	BIC	AIC	AICc	MML_u	BIC^*	AIC^*	AIC_c^*	MML_u^*
(NNG, KL, EIC)	**1**	3/200	10	10	0	0	0	0	8	10	8	9
	2	0/200	2	2	0	0	0	0	0	0	0	0
	3	0/200	1	1	0	0	0	0	0	0	0	0
	4	0/200	1	1	0	0	0	0	0	0	0	0
	5	0/200	1	1	0	0	0	0	0	0	0	0
	6	0/200	1	2	0	0	0	0	0	0	0	0
		Overall sum of correct identifications	54	53	50	50	50	50	58	60	58	59
(NNG, L, EIC)	**1**	3/200	10	10	0	0	0	0	7	10	8	8
	2	0/200	2	3	0	0	0	0	0	4	0	0
	3	0/200	3	2	0	0	0	0	0	3	0	0
	4	0/200	2	2	0	0	0	0	0	3	0	0
	5	0/200	3	3	0	0	0	0	0	3	0	0
	6	0/200	2	3	0	0	0	0	0	3	0	0
		Overall sum of correct identifications	48	47	50	50	50	50	57	44	58	58
(LS, KL, EIC)	**1**	3/200	10	10	0	3	0	0	10	10	8	10
	2	0/200	2	3	0	0	0	0	0	0	0	0
	3	0/200	3	2	0	0	0	0	0	0	0	0
	4	0/200	2	2	0	0	0	0	0	0	0	0
	5	0/200	3	3	0	0	0	0	0	0	0	0
	6	0/200	2	3	0	0	0	0	0	0	0	0
		Overall sum of correct identifications	48	47	50	53	50	50	60	60	58	60
(LS, L, EIC)	**1**	3/200	10	10	2	6	0	0	10	10	8	10
	2	0/200	2	3	0	1	0	0	1	2	0	0
	3	0/200	3	2	0	0	0	0	0	1	0	0
	4	0/200	2	2	0	0	0	0	0	1	0	0
	5	0/200	3	3	0	1	0	0	1	1	0	0
	6	0/200	2	3	0	1	0	0	0	2	0	0
		Overall sum of correct identifications	48	47	52	53	50	50	58	53	58	60

In summary of all the above examples, we find:

- Method (NNG, KL, IC), (LS, KL, IC), (LS, L, IC), (NNG, KL, EIC) and (LS, KL, EIC) worked generally well for Example 1. Among them, (NNG, KL, IC) worked the best;
- Method (NNG, KL, IC), (LS, KL, IC) and (LS, L, IC) performed well for Example 2 with (NNG, KL, IC) being the best;

- Method (NNG, KL, EIC), (NNG, L, EIC), (LS, KL, EIC) and (LS, L, EIC) performed better in Example 3 in terms of individual selection. Among them, method (LS, KL, EIC) worked best for both individual selection and group selection.

Therefore, we will pick (NNG, KL, IC) from Example 1 and 2, and (LS, KL, EIC) from Example 3, and apply two methods to real data in the next section.

5.6 Real Data Experiments

5.6.1 Real data – Birth Weight Dataset

We compared our proposed group DSS procedure to the BSGS method using the birth weight dataset from [160]. The data was collected by the Baystate Medical Center, Springfield, Massachusetts, during 1986. There were $n = 189$ observations and $p = 8$ predictors. The response variable was birth weight measured in grams. The eight predictors included mother's age (AGE), weight in pounds at the last menstrual period (LWT), race (RACE: white, black and others), smoking status during pregancy (SMOKE: yes or no), history of premature labor (PTL: 0, 1, 2, etc.), history of hypertension (HT: yes or no), presence of uterine irritability (UI: yes or no), and number of physician visits during the first trimester (FTV: 0, 1, 2, etc.). We adopted the same procedure as [37] and expanded the two continuous variables AGE and LWT using third-order polynomials treating each as a separate group. The categorical variable RACE contained three factors and was turned into two dummy variables, which also formed a group. The rest of the variables were unaltered and formed separate singleton groups. We used our grouped DSS procedure with the grouped horseshoe prior to select several sparse models using the four information criteria discussed in Section 5.3.2. The models selected by all four information criteria using the posterior expected degrees of freedom (5.34) using (NNG, KL, IC) were identical. The final model included five

groups

$$\begin{aligned}\mathbb{E}(Y|X) =& \text{LWT} + \text{LWT}^2 + \text{LWT}^3 + \text{RACE}_{\text{black-white}} + \text{RACE}_{\text{others-white}} \\ &+ \text{SMOKE} + \text{HT} + \text{UI}.\end{aligned} \tag{5.38}$$

If we use (LS, KL, EIC) method, four criteria BIC*, AIC* , AICc*, MMLu* selected 1, 5, 5, 8 groups respectively. The BIC* selected the model

$$\mathbb{E}(Y|X) = \text{UI}, \tag{5.39}$$

AIC* , AICc* selected the same model as before which is shown in 5.38 and the MMLu* selected all groups. This is probably due to the extended penalty in the information criteria which penalises less heavily when selecting either the full model or the null model.

This same dataset was previously analysed by [161] whose final model included LWT, LWT^2, RACE, SMOKE, HT, UI, and HT:RACE, where HT:RACE is an interaction term. In the original paper of [77], interaction terms were not included in the analysis; to make our experiments comparable to the experiment in the [77], we omitted all interaction terms in our analysis. The BSGS method was also used to investigate the same dataset and selected six groups: LWT, LWT^2, LWT^3, $\text{RACE}_{\text{black-white}}$, $\text{RACE}_{\text{others-white}}$, SMOKE, HT, UI, and PTL, with the posterior probability of PTL (0.512) barely passing the threshold value 0.5.

We also compared our method with the BSGS procedure in terms of predictive performance. We divided the dataset into two parts, a training dataset (75% of samples) and a testing dataset (25% of samples), and ran 100 simulations using the two algorithms. The training data were used to find a sparse estimate and the testing data were then used to compute square prediction errors for the estimates obtained from the training data. For the BSGS method, the best τ was selected from the set $\{1.5, 2, 2.5, 3\}$ for each of the training datasets.

The ratios of the mean-squared prediction errors for each method using both (NNG, KL, IC) and (LS, KL, EIC), all relative to the mean-squared prediction error achieved by the BSGS procedure, are shown in Table 5.16. The posterior mean estimates $\bar{\boldsymbol{\beta}}$ produced the best prediction error of all the methods tested; however, the posterior mean estimate is never sparse. The prediction errors obtained using the variance explained and excess error criteria are essentially identical to those obtained by BSGS. In contrast, our proposed criteria performed better than the BSGS, variance explained and excess error procedures. Using (LS, KL, EIC) method, the BIC* and MMLU* provided an mean-squared error larger than BSGS method.

TABLE 5.16: The ratio of mean-squared prediction errors for each method relative to the BSGS procedure for the birth weight data.

Method	VarExp	ExcErr	BIC*	AIC*	AIC$_c^*$	MML$_u^*$	$\bar{\boldsymbol{\beta}}$
(NNG, KL, IC)	1.002	0.984	0.979	0.942	0.946	0.940	0.906
(LS, KL, EIC)	1.002	0.984	1.005	0.973	0.976	1.007	0.906

5.6.2 Real data – Drug-Gene Interaction Study

In this section, we applied our proposed method to a drug-gene study which also used in [77]. The gene expression profiles are the predictors and the drug sensitivity profiles are the response variables.

The gene expression data and drug sensitivity profiles can be downloaded from Developmental therapeutics Program NCI/NIH website https://tinyurl.com/ybbw3scm. Both expression data and drug sensitivity profiles were performed on the same NCI-60 panel of human tumour cell lines. The micro-array gene chips were generated by Gene Logic using Affymetrix U133 platform and there are 44,693 probes for each chip. The drug response profiles from a short list of 118 drugs with known mechanism of actions were included in our analysis as the response variable. The pathway information is from the Kyoto Encyclopedia of Genes and Genomes (KEGG) database [162–164]. A total of 212 pathways and 9973 probes are included in the analysis, with each pathway containing multiple

genes and with most genes associated with multiple pathways. Chen et al. [77] further reduced the dimension of data by choosing 1000 probes with the largest coefficients of variances.

We applied our method to one of the drugs Bortezomib as in [77]. Bortezomib is an anti-cancer drug whose known mechanism of action is proteasome inhibition. There are $n = 59$ samples and $p = 1000$ predictors in this example. The group structure is defined using both gene and pathway information. There are two levels of group structures: probe-gene level and gene-pathway level. In the first level, one probe is grouped to one gene, and some of the probes may share the same gene and some of them may not belong to any known genes. We treat probes with unknown gene information as separate groups. The second level is the gene-pathway level, where each gene may belong to multiple pathways and gene-pathway information can be overlapping.

In the Bayesian sampling procedure, we took both probe-gene information and gene-pathway information into our model because our model allows for multilevel and overlapping group structures. Our aim is to select a group of genes that are associated with the Bortezomib, and therefore, in the following group NNG step, we only considered the group level with probe-gene information. We used the grouped Bayesian horseshoe model, discarded first $10,000$ burn-in samples, picked a thinning of 5, and obtained $100,000$ posterior samples. The mean of posterior samples along with the gene-pathway information were used next in the group NNG step to obtain a solution path for total 924 groups. Finally, the groups are selected using information criteria.

The numbers of groups/genes each method selected using both (NNG, KL, IC) and (LS, KL, EIC) were included in the Table 5.17. The BIC, AICc and MMLu methods selected identical 32 groups. For those 32 genes, there are 34 probes in total. Compared to the results from [77], they selected 109 probes with 24 probes involving in drug metabolism pathways, and 4 genes (*CUL4A*, *UBE2G1*, *UBE2G2*, *CBLC*) involving in ubiquitin-proteasome pathway which is the direct

targets of Bortezomib. For 34 probes we selected, 13 of them were involved in drug metabolism pathways, and 3 genes (*CUL4A*, *UBE2G1*, *UBE2G2*) are selected.

Results of using (LS, KL, EIC) approach suggested that BIC* and MMLu* selected 0 genes. The simulation results for Example 3 in Section 5.5.3 suggests that (LS, KL, EIC) performed quite well in both individual selection and group selection, and both VarExp and ExcErr are prone to overfitting. Therefore we could believe that there may be no signal in this gene data.

TABLE 5.17: The number of groups/genes each method selected for the Bortezomib using (NNG, KL, IC) and (LS, KL, EIC).

Method	VarExp	ExcErr	BIC*	AIC*	AIC_c^*	MML_u^*
(NNG, KL, IC)	22	46	32	33	32	32
(LS, KL, EIC)	22	46	0	32	25	0

5.7 Conclusion

In this chapter, we proposed a group Bayesian global-local shrinkage hierarchical model which incorporates overlapping and multilevel group structures. We extended two shrinkage priors, the horseshoe and the horseshoe+ priors, to the setting of grouped variables. We also adapted the decoupled shrinkage and selection (DSS) method to handle grouped regression models, and use the grouped non-negative garrotte to sparsify posterior estimates. To select a final sparse model, we combined four information criteria approaches with the grouped DSS procedure and proposed an improved estimate of the degrees of freedom for a candidate model in the grouped non-negative garotte solution path. As discussed in Section 5.3.1, our group DSS procedure can be extended to select groups for multilevels, however, the detail is beyond the scope of this paper and will be considered in future work.

Simulation results showed that the procedure using our proposed estimate of the degrees of freedom performed well in selecting active groups, especially when these groups were sparse. Our proposed method demonstrated similar performance to

the Bayesian grouped slab-and-spike (BSGS) method in terms of the overall rate of correct group identifications, while requiring substantially less computational resources. The performance of our proposed method in terms of group selection for Example 3 shows a good results as both BIC* and MMLu* select the active group all the time and selected no false groups. However, the performance of individual selection under the small-n-large-p situation is not as good as BSGS method. Although our proposed method selected true predictors all the time, it tends to pick more false positives compared to BSGS method.

In all simulations, the use of small-sample information criteria to select sparse models appeared to be preferable to the original heuristic model selection criteria proposed in [14]. (see Section 5.4)

Since our proposed method tends to overfit inactive groups in small-n-large-p scenario, we extended our proposed method to allow for three variations in the approach: (1) use different estimates of the coefficients; (2) utilise the exact likelihood in the information criteria, and (3) use additional penalty terms in the likelihood (see Section 5.5). Then, we applied those eight combinations to all three examples. Simulation results of all eight combinations showed that most methods worked well for Example 1 where is moderate size of active groups. The EIC-type methods did not work well for Example 2 because they tend to overfit the inactive groups. For Example 3, only (NNG, KL, EIC) and (LS, KL, EIC) worked well in terms of individual selection since they did not over select the inactive groups. Both of them also performed well in group selection since they selected active groups all the time and selected no false groups.

We applied both (NNG, KL, IC) and (LS, KL, EIC) to real data sets: birthweight dataset and gene-drug dataset. The prediction errors of our proposed method (NNG, KL, IC) obtained in an experiment using real data also showed an improvement over the BSGS method and the original DSS model selection criteria. These results suggest that the proposed grouped DSS procedure with information criteria is an efficient tool for selecting sparse grouped models with good predictive performance.

We would recommend to use (NNG, KL, IC) to perform all group selections, and (LS, KL, EIC) to do individual selection under small-n-large-p situation.

Chapter 6

Application of Bayesian Grouped Model to Breast Cancer Data

6.1 Introduction

Breast cancer is one of the most common cancer for women, and genetic is an important etiology for breast cancer. GWAS, as introduced in Section 3.1, have particularly focused on identifying common genetic variants associated with high effect size or severe disease. The conventional method for GWAS is the SNP–by–SNP univariate testing which has been discussed in Section 3.1.3. Although the marginal testing can be easily implemented for a large dataset, it has several limitations (see Section 3.1.3).

In this chapter, we introduce a new gene-based testing motivated by the fact that a gene consists of many different SNPs which contribute independently to a disease. In addition, population stratification is another issue and it can result in potential spurious association in GWAS [89]. Principal component (PC) analysis [89] has been applied to the data to account for the systematic ancestry differences among sample individuals. Therefore, predictors of PCs and SNPs should be considered as two separate groups. We propose to apply our Bayesian grouped model with a global-local shrinkage prior, the horseshoe prior in Section 5.2.3, to each gene along

with the PCs. The Bayesian horseshoe model fits groups of SNPs that belong to the same gene and PCs simultaneously, and it has been proved to promotes sparsity.

After we obtain posterior samples from Bayesian model, we apply posterior likelihood ratio (PLR) method to posterior samples. Compared with conventional likelihood ratio test, the posterior likelihood ratio method can be easily used in conjunction with sparsity promoting global-local shrinkage prior from Bayesian grouped model for handling sparsity. It is particular useful for our GWAS where only a small amount of SNPs may be associated with the disease, and it can be used simply to compute p-values for all features.

We then propose to apply the gene-based method to a real GWAS breast cancer data, Haiman-Hopper (H2) dataset, to find significant genes or SNPs that are associated with breast cancer. We also build a predictive model that can be employed for foreseeing patients with future risk of getting the disease.

The organisation of this chapter is:

- The conventional approach to analyse GWAS is marginal testing (see Section 3.1.3). We first introduce the classic approaches to hypothesis testing: the likelihood ratio method and Bayes factor. We then introduce the posterior likelihood ratio method which can be easily applied in conjunction with sparsity construction of our Bayesian model.

- We compare the performance of the PLR with the conventional method LR, and results of PLR from simulation studies are promising. Therefore, we propose the gene-based testing method by combining our Bayesian grouped regression methodology introduced in Chapter 5 with PLR.

- Next, we introduce a real GWAS dataset, Haiman-Hopper dataset and apply the proposed gene-based testing method to it. We fit a generalised Bayesian horseshoe model to the dataset to obtain a set of posterior samples, then using the posterior likelihood ratio method to compute p-values for each gene.

- After that, we compare the findings from our method with existing literature in breast cancer studies. We also compare the results of our selected genes or SNPs with two existing gene-based testing methods.
- Finally, we build a predictive model for H2 dataset and compute an area under the curve (AUC) of the model.

6.2 Hypothesis Testing

In this section, we will introduce some general approaches to hypothesis testing. Hypothesis testing has been widely used in genome-wide association studies. As introduced in Section 3.1.3.1, the main idea is to apply hypothesis testing to each individual SNPs and test for their p-values.

More formally, consider testing a simple point null hypothesis against a composite alternative. The hypothesis testing is to test the null hypothesis is $H_0 : \theta = \theta_0$ versus the alternative hypothesis $H_1 : \theta \neq \theta_0$.

6.2.1 Likelihood Ratio Test

One of the most common way of testing statistics for a null model against an alternative model is the likelihood ratio (LR) test which is based on the likelihood ratio of two models: the null model and the alternative model.

The test statistic of LR is usually denoted as D, and it is twice the difference in logarithm of the likelihood ratios between the null model and the alternative model. It is expressed as

$$\begin{aligned} D &= -2\ln\left(\frac{\text{likelihood for null model}}{\text{likelihood for alternative model}}\right) \\ &= 2\ln\left(\frac{\text{likelihood for alternative model}}{\text{likelihood for null model}}\right) \\ &= 2\ln\left(\frac{L(\theta_1)}{L(\theta_0)}\right) \end{aligned}$$

The probability distribution of the test statistic D is approximately a chi-squared distribution of degrees of freedom equal to $df_{\text{alt}} - df_{\text{null}}$ where df_{alt} is the number of free parameters in the alternative model and df_{null} is the number of free parameters in the null model. The p-value of the test can therefore be derived, and the null hypothesis is rejected at α level if p-value is smaller than $\alpha\%$ which implies that the alternative fit is significantly better so that it is favored.

6.2.2 Bayes Factor

A Bayesian alternative to classical hypothesis testing is the Bayes factor (BF) [16, 19, 165],. The BF is a statistical index that quantifies the evidence for one hypothesis, compared to another hypothesis. The Bayes factor, represents how our relative beliefs should change in terms of the data, and it is therefore, the relative evidence in the data. It is the evidence in the data that one hypothesis is favored over another to the same degree that one hypothesis predict the observed data more desirably than another.

The prior odds describe the degree to which one hypothesis is favored over another before the data is observed, and these can be expressed as:

$$\frac{P(H_1)}{P(H_0)}.$$

The posterior odds describes the degree to which one hypothesis is favored over another after the data is observed, and it can be calculated using Bayes' theorem:

$$\frac{P(H_1|y)}{P(H_0|y)} = \frac{P(y|H_1)}{P(y|H_0)} \cdot \frac{P(H_1)}{P(H_0)},$$

where y is the observed data. The Bayes factor is the ratio of the likelihood probability of two hypotheses, H_1 and H_0, and it can be expressed as

$$\text{BF}_{10} = \frac{P(y|H_1)}{P(y|H_0)} = \frac{\int P(\theta_1|H_1)P(y|\theta_1, H_1)d\theta_1}{\int P(\theta_0|H_0)P(y|\theta_0, H_0)d\theta_0} = \frac{P(H_1|y)P(H_0)}{P(H_0|y)P(H_1)}, \tag{6.1}$$

which shows that the BF is calculated by integrating the likelihood with respect to the priors on parameters.

However, one potential limitation for the BF is the evaluation of the integral in (6.1). The integral term

$$\int P(\theta_1|H_1)P(y|\theta_1, H_1)d\theta_1$$

needs to be computed analytically and it is often intractable, which makes it difficult to compute the BF.

6.2.3 Posterior Likelihood Ratio

In this section, we will introdcue the posterior likelihood ratio method. Let $L(\theta)$ be the likelihood function of the model $p(\mathbf{y}|\theta)$ for a given data y, and let $\pi(\theta)$ be the prior distribution for θ.

The posterior distribution of the likelihood ratio $L(\theta_0)/L(\theta)$ was first considered by Dempster [166, 167], and was then extended by Aitkin [168]. The likelihood ratio $\mathrm{LR}(\theta)$ of the null hypothesis $H_0 : \theta = \theta_0$ versus the alternative hypothesis $H_0 : \theta \neq \theta_0$ was proposed, and the null hypothesis is rejected if

$$\mathrm{LR}(\theta_0) = \frac{L(\theta_0)}{L(\theta)} < k, \tag{6.2}$$

where θ is an unknown parameter under the alternative hypothesis and $k > 0$ is a constant that is required to be chosen. If k is chosen to be 1, then it implies that H_1 is more convincing than H_0. The above equation (6.2) is equivalent to

$$l(\theta_0) - l(\theta) < \log(k),$$

where $l(\cdot)$ is the log-likelihood function.

Although θ in the alternative hypothesis is unknown, $L(\theta)$ is a function of θ. So the posterior distribution given the data $\pi(L(\theta)|\mathbf{y})$ can be computed from the

posterior distribution of θ and the posterior distribution of $\pi(\theta|\mathbf{y})$. The posterior distribution of the likelihood ratio $\pi(\mathrm{LR}(\theta)|\mathbf{y})$ can thus be computed given θ_0 is known.

Therefore, the amount of certainty in the strength of evidence against the null hypothesis H_0 can be measured by the pair (k, π_k), where π_k is the posterior probability of LR less than k, and it is mathematically defined as

$$\begin{aligned}\pi_k &= Pr\left(\mathrm{LR}(\theta) < k|\mathbf{y}\right) \\ &= Pr\left(l(\theta_0) - l(\theta) < \log(k)\,|\,\mathbf{y}\right).\end{aligned}$$

If k decreases, then π_k also reduces, and thus a smaller posterior probability π_k of strength is required to achieve a stronger evidence against H_0. So it can used to evaluared the strength in the evidence of one hypothesis versus another.

Posterior likelihood ratio method can be easily used in conjunction with sparsity promoting global-local shrinkage prior for handling sparsity which is suitable for our GWAS where only a small amount of SNPs may be associated with the disease, and it can be used simply to compute p-values for all features. Therefore, we propose the use of the PLR on GWAS data to find significant genes or SNPs. Additionally, we will compare the performance of the PLR with the conventional method LR.

6.3 Simulation Results for Posterior Likelihood Ratio Method

Before we apply posterior likelihood ratio method to real data, we first examine the performance of PLR in a simulation study. To test the performance of PLR, we first run some experiments on the simulated data for calculating type I and type II errors. Type I error is the incorrect rejection of a true null hypothesis (i.e. false discovery) and type II error is the failure to reject a false null hypothesis.

Recall in 5.1, the likelihood of a general hierarchical Bayesian global-local shrinkage regression model can be expressed as

$$\mathbf{y} \,|\, \mathbf{X}, \boldsymbol{\beta}, \sigma^2 \;\sim\; N(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I}_n),$$

which means that the response variable $\mathbf{y}$ is generated based on a normal distribution with mean $\mathbf{X}\boldsymbol{\beta}$ and variance σ^2.

To compute test statistics for PLR and LR, both the null model and the alternative model are fitted to the data separately and the log-likelihood values are recorded. Suppose we use $\mathbf{X}_{\text{null}}$ for the null model, where $\mathbf{X}_{\text{null}}$ is simply the first $M(< p)$ columns of $\mathbf{X}$, and we use $\mathbf{X}$ for the alternative model. For the null hypothesis we set the coefficients for first M columns to be non-zero, while the rest are all zeros, and it can be represented as

$$H_0 : \beta_1 \neq 0, \ldots, \beta_M \neq 0, \beta_{M+1} = \ldots = \beta_p = 0,$$

and the alternative hypothesis we set to test is that first M columns are non zeros, and at least one of the remaining coefficients are non-zero. In this case, we are able to test if any of the remaining predictors are significant or not. The alternative hypothesis can be represented as

$$H_1 : \beta_1 \neq 0, \ldots, \beta_M \neq 0, \text{ and } \beta_j \neq 0, j \in \{M+1, \ldots, p\}.$$

To simplify the problem in testing the performance, we let $M = 1$, which means only the first column of $\mathbf{X}$ is our null model. So when computing type I errors, the coefficient was set to be $\boldsymbol{\beta} = (1, 0, \ldots, 0)'$. For computing type II errors, the coefficient was generated using either sparse model $\boldsymbol{\beta} = (1, \beta_2, 0, \ldots, 0)'$ or dense model $\boldsymbol{\beta} = (1, \beta_2, \ldots, \beta_p)'$, where β_j are randomly generated from a standard normal distribution. We then generated $t = 1000$ data sets including $p = 10$ candidate predictors $\mathbf{X}_{n \times p}$ from the multivariate normal distribution with correlation structure being either (i) no correlation or (ii) pairwise correlation of 0.5. Here, pairwise correlation of 0.5 means that $\text{corr}(X_i, X_j) = 0.5, \forall i \neq j, i, j \in \{1, \ldots, p\}$.

We varied $n = \{20, 100\}$, and we also generated σ^2 according to signal-to-noise ratio SNR= $\{1, 3, 8\}$. For each realisation, we generated $\mathbf{y}$ based on a normal distribution with mean $\mathbf{X}\boldsymbol{\beta}$ and variance σ^2. Then we fitted a Bayesian model twice to both null model and the alternative model to obtain the log-likelihood functions. The prior distribution we varied is by using either a ridge or a horseshoe prior.

To compare the performance, the conventional LR method was also applied to t samples of the data. We obtained the type I errors and type II errors for each method, and based on type II errors we also computed the power of each method. The results of type I error and power of both methods are shown in Table 6.1 and Table 6.2 respectively. From the table, we see that the PLR has a much lower type I error rate compared to the LR under all combinations of different settings. The power of PLR performed worse compared to the power of LR for almost all scenarios especially for the horseshoe model with SNR = 1 and a dense coefficient. However, as both sample size and SNR increase, the power of PLR converges to 1. This is because LR is not conservative, and it tends to always include more predictors and therefore, select the alternative model more often than PLR does. For all settings with larger SNR ratio and larger samples, such as $n = 100$ and SNR $= \{5, 8\}$, the power of PLR behaved well and it was almost always equal to 1. Therefore, we expect the PLR method to produce less false positives than the LR method.

TABLE 6.1: Type I error of both PLR and LR method for horseshoe (hs) model and ridge model when correlation structure is either no correlation or pairwise correlation with $\rho = 0.5$, sample size $n = \{20, 100\}$ and signal-to-noise ratio SNR $= \{1, 3, 8\}$.

Correlation	Model	n	SNR	PLR	LR	Model	n	SNR	PLR	LR
No Correlation	hs	20	1	0.021	0.249	ridge	20	1	0.013	0.256
		20	3	0.045	0.263		20	3	0.016	0.257
		20	8	0.056	0.304		20	8	0.015	0.265
		100	1	0.024	0.064		100	1	0.034	0.085
		100	3	0.030	0.087		100	3	0.020	0.063
		100	8	0.040	0.071		100	8	0.022	0.062
Pairwise Correlation $\rho = 0.5$	hs	20	1	0.03	0.304	ridge	20	1	0.014	0.271
		20	3	0.035	0.282		20	3	0.02	0.275
		20	8	0.034	0.266		20	8	0.025	0.282
		100	1	0.022	0.063		100	1	0.025	0.071
		100	3	0.029	0.075		100	3	0.03	0.074
		100	8	0.025	0.068		100	8	0.025	0.075

TABLE 6.2: Power of both PLR and LR method for horseshoe (hs) model and ridge model when coefficient $\boldsymbol{\beta}$ is either sparse or dense, correlation structure of $\mathbf{X}$ is either no correlation or pairwise correlation with $\rho = 0.5$, sample size $n = \{20, 100\}$ and signal-to-noise ratio SNR $= \{1, 3, 5, 8\}$.

$\boldsymbol{\beta}$	Correlation	Model	n	SNR	PLR	LR	Model	n	SNR	PLR	LR
Sparse	No Correlation	hs	20	1	0.05	0.37	ridge	20	1	0.048	0.451
			20	3	0.935	0.98		20	3	0.026	0.383
			20	5	1	1		20	5	0.14	0.754
			20	8	1	1		20	8	0.736	0.998
			100	1	0.830	0.924		100	1	0.432	0.635
			100	3	1	1		100	3	1	1
			100	5	1	1		100	5	1	1
			100	8	1	1		100	8	1	1
	Pairwise Correlation with $\rho = 0.5$	hs	20	1	0.085	0.454	ridge	20	1	0.028	0.310
			20	3	0.028	0.278		20	3	0.056	0.450
			20	5	1	1		20	5	0.984	1
			20	8	1	1		20	8	0.985	1
			100	1	0.248	0.357		100	1	0.320	0.450
			100	3	1	1		100	3	0.999	1
			100	5	1	1		100	5	1	1
			100	8	1	1		100	8	1	1
Dense	No Correlation	hs	20	1	0.059	0.378	ridge	20	1	0.053	0.484
			20	3	0.439	0.977		20	3	0.233	0.84
			20	5	0.933	1		20	5	0.224	0.893
			20	8	0.991	1		20	8	1	1
			100	1	0.103	0.216		100	1	0.128	0.37
			100	3	1	1		100	3	0.996	0.999
			100	5	1	1		100	5	1	1
			100	8	1	1		100	8	1	1
	Pairwise Correlation with $\rho = 0.5$	hs	20	1	0.031	0.371	ridge	20	1	0.017	0.483
			20	3	0.742	0.999		20	3	0.578	0.985
			20	5	0.098	0.364		20	5	0.97	1
			20	8	0.95	1		20	8	1	1
			100	1	0.244	0.426		100	1	0.516	0.839
			100	3	0.794	0.871		100	3	1	1
			100	5	1	1		100	5	1	1
			100	8	1	1		100	8	1	1

6.4 Haiman-Hopper dataset

Genome-wide association study of breast cancer in high-risk women is a new genome-wide association study which aims to identify novel genetic variants associated with breast cancer in high-risk women. The dataset for analysis is the Haiman-Hopper (H2) dataset of breast cancer. To date, five million SNPs have been genotyped for over 3,000 breast cancer European ancestry cases with a strong family history of the disease, and for over 3,000 European ancestry controls. To be more specific, there are $n = 6955$ observations with 3515 cases and 3440 controls in the dataset. Genotypes are encoded using additive model as discussed in Section 3.1.2.2. We divided the dataset into 22 subsets with each dataset corresponding to one chromosome. The basic information of the dataset is summarised in Table 6.3. For each chromosome, we display the total number of genes, average number of SNPs in each gene, minimum number of SNPs in each gene and maximum number of SNPs in each gene.

6.5 Gene-Based Testing for H2 Dataset

In this section, we will apply gene-based method to H2 dataset to test and see if any gene is associated with breast cancer. We will first introduce the procedure for gene-based testing based on our Bayesian grouped regression methodology introduced in Chapter 5, and then we will run the experiment and show the results.

6.5.1 The Procedure for Gene-Based Testing

As mentioned in Section 6.4, we have 22 subsets where each subset corresponds to a chromosome. Within each chromosome, we looked at each gene separately by fitting a Bayesian horseshoe model, obtaining posterior samples from the Bayesian model, and then performing posterior likelihood ratio tests to compute p-values for each gene as described in Section 6.2.3.

TABLE 6.3: Basic information of the H2 dataset: chromosome number, total number of genes in each chromosome, mean number of SNPs in each gene, minimum number of SNPs in each gene within the chromosome and maximum number of SNPs in each gene within the chromosome.

Chromosome number	Total number of genes	Mean number of SNPs	Minimum number of SNPs	Maximum number of SNPs
1	2601	76.7	1	1364
2	1705	96.9	1	2310
3	1484	99.6	1	1801
4	1001	104.6	2	1460
5	1223	97.4	1	1195
6	1390	101.3	3	1819
7	1243	93.7	1	2561
8	964	105.5	1	6158
9	1029	87.5	1	1299
10	1045	102.2	1	1822
11	1621	76.5	1	1737
12	1326	85.0	3	1434
13	600	109.0	11	1387
14	873	85.1	2	1817
15	907	83.4	1	1319
16	1041	80.9	1	3945
17	1494	67.8	1	1016
18	407	120.5	1	1465
19	1720	57.8	2	447
20	752	84.81	2	2409
21	355	91.8	1	1291
22	595	78.2	1	947

For H2 dataset, there is one variable called region which consists of 9 regions: Australia, Canada, Germany, Hawaii/Los Angeles/SF, Midwest, North East coast (NY/Phil), Spain, UK and Utah, so it can be represented using 8 dummy variables.

There are different subgroups in the population so that the data may involve ancestry differences. Without adjusting for ancestry differences, it may lead to spurious associations. A common approach to detect and correct for population stratification is principal component analysis [89, 169, 170].

The top principle components are obtained by applying principal components analysis to the SNPs matrix and therefore the dimensionality of the model is reduced. Then it uses the top eigenvectors of the sample covariance matrix as covariates in a multi-linear regression model because a few top PCs usually explain most of the variation in the data. Finally it computes association statistics for

hypothesis testing. For H2 dataset, the top PCs are selected using EIGENSTRAT method [89] and there are 9 principal components (PCs) obtained.

To apply PLR, we need to specify the null model and the alternative model. The null model includes 17 predictors: 9 PCs which are extracted from the data and 8 dummy variables from the region variable. For gene-based testing, the alternative model for each gene has 17 predictors which are the same as those from the null model, plus all SNPs within that gene. The detailed procedure for selecting top genes using gene-based testing is:

1. Fit a Bayesian logistic model to the 17 predictors and obtain likelihood statistics for the null model.
2. For each gene, fit a group Bayesian logistic model with horseshoe prior to all SNPs plus 17 predictors where we treat all SNPs within a gene as a group and all 17 predictors as another group.
3. Use the PLR method to compute posterior likelihood ratio (individual gene model divided by null model) to get a series of p-values for each gene.
4. Compute statistics for LR method and obtain p-values for each gene.

6.5.2 Results of Selected Genes for H2 Dataset

After we performed gene-based testing by fitting both a Bayesian logistic model with horseshoe prior to 17 predictors only and a Bayesian group logistic model with horseshoe prior to each individual gene plus 17 PCs to obtain statistics for both null model and alternative model, both with 10,000 samples, we then computed p-values for all genes.

The results of p-values for all genes are displayed as a Manhattan plot as shown in Figure 6.1, where the y-axis is the $-\log_{10}$ transformed p-values that indicate the power of association and the x-axis are all chromosomes. The plot indicates the significance of the association between a gene and breast cancer. The higher

the point is located on the graph, the lower the p-value is for the gene, and therefore, the more significant the gene is associated with the disease. One gene in chromosome 10 has the highest $-\log_{10}$ value, which implies that it has the largest association with breast cancer according to our model. From the plot, we see that genes in chromosome 3, 8, 9, 10, 11, 16 on average have larger $-\log_{10}$ values and therefore higher degree of association with breast cancer.

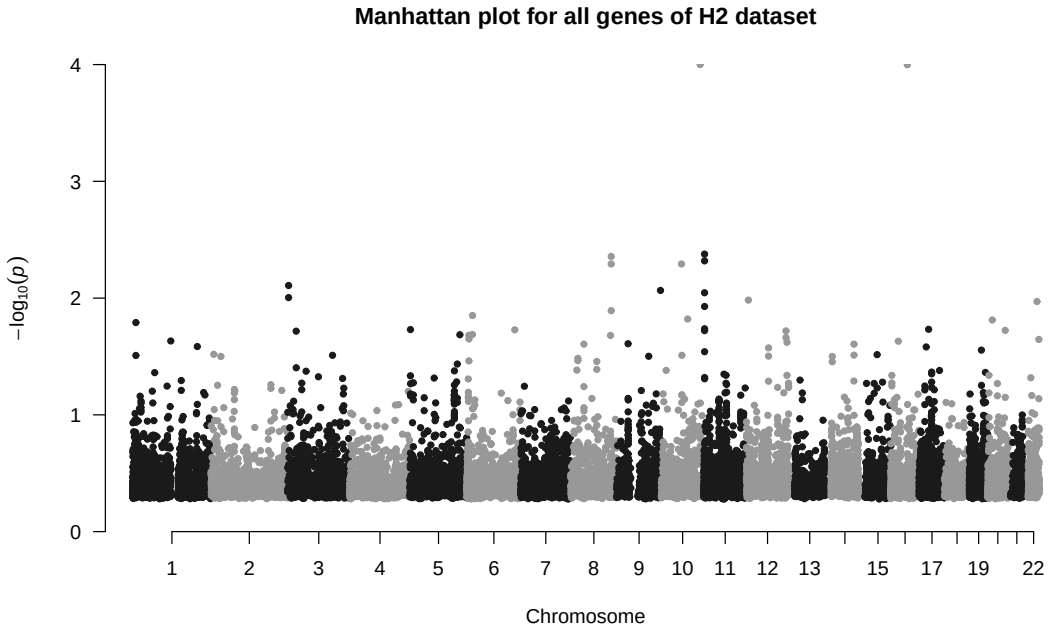

FIGURE 6.1: Manhattan plot for all genes in 22 chromosomes of Haiman-Hopper dataset.

There were 86 genes with p-values smaller than 5%. Table 6.4 and Table 6.5 show the list of 86 genes. Since we used 10,000 samples in the model, if a gene appears only once out of 10,000 samples, then the p-value is $1/10,000 = 0.0001$; if a gene appears zero time out of 10,000 samples, then the p-values is 0. Therefore, the smallest nonzero p-values we can observe is 0.0001. Top 3 genes: *FGFR2*, *CASC16* and *TOX3*, all appeared 0 times out of total 10,000 samples in the model, so p-values of those genes should be regarded as smaller than 0.0001.

TABLE 6.4: The information of the list of 86 genes selected.

Rank	Chr number	No. in chr	Gene name	Number of SNPs	p-values of PLR
1	10	312	*FGFR2*	164	< 0.0001
2	16	107	*CASC16*	95	< 0.0001
3	16	936	*TOX3*	161	< 0.0001
4	11	695	*LSP1*	104	0.0042
5	8	90	*CASC8*	352	0.0044
6	11	556	*LINC01150*	52	0.0048
7	8	89	*CASC21*	263	0.0051
8	10	1030	*ZNF365*	157	0.0051
9	3	1379	*TRNT1*	133	0.0078
10	9	299	*GLT6D1*	98	0.0086
11	11	753	*MIR4298*	60	0.009
12	3	508	*IL5RA*	238	0.0099
13	12	1293	*VWF*	267	0.0104
14	22	252	*LOC101927526*	83	0.0107
15	11	1506	*TNNT3*	59	0.0118
16	8	94	*CCAT2*	69	0.0128
17	6	910	*NUP153*	100	0.0141
18	10	713	*NRG3*	1204	0.0151
19	20	356	*MACROD2*	2409	0.0154
20	1	1775	*PEX14*	175	0.0162
21	11	1505	*TNNI2*	60	0.0183
22	17	1085	*RAB11FIP4*	221	0.0185
23	5	715	*MIR4457*	64	0.0186
24	6	125	*CCDC170*	197	0.0187
25	20	365	*MC3R*	75	0.0189
26	11	1439	*SYT8*	58	0.019
27	12	274	*DHX37*	136	0.0191
28	3	924	*NEK10*	215	0.0192
29	6	269	*FAM8A1*	43	0.0205
30	5	227	*EBF1*	376	0.0206
31	6	825	*MIR5683*	112	0.0208
32	8	648	*NSMCE2*	239	0.0209
33	12	6	*AACS*	121	0.0218
34	6	845	*MIR7853*	112	0.0224
35	22	114	*DENND6B*	46	0.0226
36	1	543	*EMBP1*	33	0.0233
37	16	156	*CHP2*	42	0.0234
38	12	629	*LOC101927694*	79	0.0239
39	9	815	*RUSC2*	74	0.0246
40	8	23	*ANK1*	258	0.0248
41	14	830	*TRIP11*	92	0.0248
42	1	1249	*LOC284578*	51	0.026
43	17	903	*MTRNR2L1*	43	0.0262

TABLE 6.5: The information of the list of 86 genes selected (continued).

Rank	Chr number	No. in chr	Gene name	Number of SNPs	p-values of PLR
44	12	633	*LOC101928002*	65	0.0267
45	19	1222	*SIPA1L3*	293	0.0279
46	11	819	*MIR7847*	54	0.0288
47	2	833	*LOC101929567*	95	0.0303
48	15	376	*LOC101928694*	225	0.0305
49	14	53	*ATXN3*	79	0.0308
50	3	653	*LOC100507389*	91	0.0309
51	10	548	*LOC105378330*	17	0.0309
52	1	464	*DFFA*	31	0.031
53	12	99	*BEST3*	144	0.0314
54	9	278	*GABBR2*	695	0.0315
55	2	477	*GALNT14*	348	0.0316
56	14	71	*C14orf119*	59	0.0316
57	8	671	*PDLIM2*	74	0.0328
58	8	820	*SORBS3*	82	0.0343
59	6	254	*F13A1*	422	0.0345
60	8	690	*PMP2*	44	0.0349
61	14	4	*ACIN1*	58	0.0351
62	5	970	*RBM22*	33	0.0367
63	3	1212	*SLC4A7*	112	0.0395
64	8	228	*FABP9*	45	0.0408
65	8	789	*SH2D4A*	191	0.0414
66	10	281	*FAM171A1*	263	0.0416
67	17	1058	*PRKCA*	689	0.0417
68	5	875	*PCDHGA12*	39	0.042
69	3	23	*ACOX2*	66	0.0423
70	17	562	*KRTAP4-1*	37	0.043
71	19	1204	*SHANK1*	134	0.0433
72	1	1290	*LRRC7*	297	0.0435
73	17	565	*KRTAP4-2*	39	0.0445
74	11	1292	*RTN3*	48	0.0448
75	11	41	*ANO1-AS2*	76	0.0458
76	20	593	*SLC4A11*	73	0.0458
77	12	615	*LOC100996679*	155	0.0459
78	16	417	*LINC00514*	30	0.0459
79	5	1126	*TERT*	107	0.0463
80	3	962	*OR5H15*	49	0.0472
81	11	857	*MRPL23*	52	0.0479
82	22	479	*SGSM1*	231	0.0481
83	5	900	*PDE8B*	309	0.0484
84	3	908	*NAALADL2*	1168	0.0488
85	11	1387	*SNORD131*	45	0.0491
86	6	446	*HLA-DPB1*	66	0.0493

The top gene we selected from our method is *FGFR2*, which stands for fibroblast growth factor receptor 2. The FGFR2 gene has been shown to be highly associated with breast cancer [171], [135], [172]. The second ranked gene is CASC16 which stands for cancer susceptibility 16. The influence of this gene on breast cancer has also been established in the research [173], [174]. The third gene *TOX3*, *TOX* high mobility group box family member 3, whose mutations are connected with an increased risk of breast cancer [174], [175], [176], [135].

The fourth gene LSP1 represented for lymphocyte-specific protein 1, is also proven to be associated with breast cancer [177], [178], [129], [179] [172]. *CASC8*, which ranked fifth in our top selected genes, is cancer susceptibility 8, and it is shown that it has a relation to breast cancer [129], [135]. The sixth to eighted ranked genes are *LINC01150*, *CASC21*, *ZNF365* respectively, and their proven associations can be found in the literature [172], [180], [181].

The associations with breast cancer for genes *MIR4298*, *IL5RA*, *VWF*, *TNNT3*, *CCAT2*, *NRG3*, *MACROD2*, that ranked 11th, 12th, 13th, 15th, 16th, 18th, 19th respectively can be found in [182], [183], [184], [172], [185], [186], [187] respectively. The 20th ranked gene PEX14, peroxisomal biogenesis factor 14, has also been proven to have strong associations with breast cancer [188], [172], [189]. The associations between genes *TNNI2* and *MIR4457* which ranked 21st and 23rd, and breast cancer can be found in the literature [172], [190] respectively. The 24th ranked gene *CCDC170*, coiled-coil domain containing 170, is also believed to have affect the risks of getting breast cancer [191], [191].

Next, we also applied LR method to H2 dataset to obtain a set of p-values for each gene, and compared correlations between p-values for PLR, LR, along with two other methods: versatile gene-based association study (VEGAS) and the generalised higher criticism (GHC) [192] methods. The correlation matrix of p-values for each gene in H2 dataset between PLR, LR, VEGAS, GHC methods is shown in Table 6.6. The correlation between PLR and LR is the most significant among four methods, and the correlation between LR and GHC is the smallest in the table. In conclusion, most methods shown moderate to strong correlation between

each other except for the correlation between LR and GHC, which implies that there are a few common genes selected by the above approaches.

TABLE 6.6: Correlation matrix of p-values for each gene in H2 dataset between PLR, LR, VEGAS, GHC methods

Method	PLR	LR	VEGAS	GHC
PLR	1			
LR	0.6613	1		
VEGAS	0.5367	0.4023	1	
GHC	0.4547	0.3142	0.5056	1

6.6 Predictive model

We have identified multiple genes that might have impact on breast cancer in Section 6.5.2. Since cancer is largely caused by the mutations in the DNA of human cells that arise during our life instead of the DNA we inherited from ancestors. The DNA consists of a large number of genes, each of which contains a large number SNPs, so those mutations can be captured by SNPs. If a large of SNPs are included in the predictive model, it can substantially increase the predictive performance. However, it could also result in overfitting. Therefore, a desired predictive model is the one that balances an overall estimation of risk and the number of predictors.

We then build a predictive model that can be employed for foreseeing patients with future risk of getting the disease. The predictive model is built based on SNPs. The procedure for building a predictive model is:

1. We select a reasonable number of genes, say K
2. To select top P SNPs, we divide the whole dataset with $n = 6955$ samples into two parts: training data (75%) and testing data (25%).
3. For training data, we fit a group Bayesian horseshoe model to all SNPs from top K genes as well as 17 PCs, where all SNPs belong to same gene form a

group; all PCs form one group. Therefore, there are total $K + 1$ groups in the model (i.e. K groups of genes + 1 group of PCs).

4. After obtaining posterior samples from the Bayesian horseshoe model, we compute the mean of posterior samples.

5. We derived a fitted model for training data using posterior mean of the samples.

6. We then evaluate the performance of predictive model by applying the predictive model to the testing data, and computing an area under the curve (AUC) score of the model.

Following the above procedure, we selected top 10 genes which include total 1579 SNPs in total. We divided whole data into training data (75%) and testing data (25%). For the training data, we fit a group Bayesian horseshoe model to those 1579 SNPs and 17 PCs, where all SNPs belong to same gene form a group; all PCs form one group. So there are total 11 groups in the alternative model (i.e. 10 groups of genes + 1 group of PC). We computed the mean of posterior samples for the alternative model from the training data, and predicted on the testing data to get an AUC score of 0.6544 in this case.

In the next step, we further selected top 1000 SNPs besides those 17 PCs by ranking those 1579 SNPs according to their posterior z score which is the absolute value of its posterior mean divided by its standard deviation based on the posterior samples, and selected top 1000 SNPs. We then ran the group Bayesian horseshoe model again for the training data, obtained a fitted model using mean of posterior samples, and predicted the performance on testing data to get an AUC score of 0.6486. The AUC score of top 1000 SNPs is slightly worse than that of top 1579 SNPs. Table A.1 in Appendix A illustrated the list of top 1000 SNPs we selected using the described method.

6.7 Conclusion

In this chapter, we introduced the conventional likelihood ratio method (LR) and the posterior likelihood ratio (PLR) method, and test the performance of LR and PLR in terms of type I and II errors. Based on type II errors, we also computed the power of each method. The results of type I error and power of both methods showed that the PLR had a much lower type I error rate compared to the LR under all combinations of different settings. In terms of the power of each method, the LR tended to always include more predictors and therefore, selected the alternative model more often than PLR. Therefore, the power of LR performed better compared to the power of PLR for almost all scenarios. However, the power or PLR converges to 1 as both sample size and SNR increase. For all settings with larger SNR ratio and larger samples, the power of PLR behaved well and it was almost always equal to 1. Therefore, the PLR method was expected to produce less false positives than the LR method (see Section 6.3).

We then introduced a real GWAS breast cancer dataset, Haiman-Hopper (H2) dataset, and developed a gene-based testing method to detect important features of H2 dataset. The proposed method worked by applying PLR to posterior samples from our proposed Bayesian grouped model to obtain p-values for all genes. The results showed that genes in chromosome 3, 8, 9, 10, 11, 16 on average had higher degree of association with breast cancer. There were total 86 genes identified by our method with p-values smaller than 5%. Top three genes selected are *FGFR2*, *CASC16* and *TOX3* which were all believed to have associations with breast cancer in the literature (see Section 6.5.2).

After that, we compared the results of proposed method applying to H2 with existing gene-based testing methods: LR, versatile gene-based association study (VEGAS) and the generalised higher criticism (GHC) in Section 6.5.2. The correlation matrix of p-values for each gene in H2 dataset between PLR, LR, VEGAS, GHC methods was calculated. The correlation between PLR and LR is the most significant among four methods, and the correlation between LR and GHC is the smallest in the table. In conclusion, most methods shown moderate to strong

correlation between each other except for the correlation between LR and GHC, which implies that there are a few common genes selected by the above approaches.

Finally, we built a predictive model based on 1579 SNPs for H2 dataset for foreseeing patients with potential risk of getting the disease. The predictive model was built by computing posterior mean from a group Bayesian horseshoe model for the training data. The AUC score of our predictive model on the testing data was 0.6544. In the next step, we further selected top 1000 SNPs besides those 17 PCs by ranking those 1579 SNPs according to their posterior z score and selected top 1000 SNPs. Following the same procedure, we obtained an AUC score of 0.6486 which was slightly worse than that of top 1579 SNPs (see Section 6.6).

Chapter 7

Conclusion and Future Work

7.1 Thesis Summary

This thesis has investigated Bayesian models for grouped variables. The Bayesian grouped horseshoe model was proposed. In addition, a more general Bayesian sparse global-local shrinkage regression model for grouped variables was proposed to handle complex group structures. The group decoupled shrinkage and selection technique together with generalised information criteria approaches with an alternative degrees of freedom estimator was proposed to perform sparse group selections. Finally, the posterior likelihood ratio method was incorporated into our proposed Bayesian group models, and was applied to a real GWAS breast cancer dataset for selection and prediction.

The Bayesian horseshoe estimator is known for its robustness when handling noisy and sparse big data problems. In Chapter 4, we have presented two extensions of the regular Bayesian horseshoe (BHS): (i) the grouped Bayesian horseshoe (BGHS) and (ii) the hierarchical Bayesian horseshoe (HBGHS). The advantages of the proposed methods are their flexibility of handling grouped variables through extra shrinkage parameters at the group and within-group levels. We then introduced non-parametric additive regression models, where each predictor is a set of basis

functions that naturally form a group structure which we tested the performance of our models on.

We applied HGBHS along with BHS, BHS with no expansion, and lasso with BIC to a set of simulated datasets that are with or without group structures. When the simulated data come from a simple linear function, the average mean squared prediction error for BHS without expansion unsurprisingly performed best; HBGHS was competitive with BHS for small sample size n. As sample size grew, HBGHS outperformed BHS. When there was a highly nonlinear relationship in the data, the HBGHS almost always achieved the smallest mean squared prediction error, except for only one scenario when sample size was small, $n = 100$, and signal-to-noise ratio was low (SNR = 1). When the underlying data consisted of polynomial expansions, HBGHS consistently gave the best performance for mean squared prediction error, followed by BHS. In summary, the HBGHS method outperformed the regular BHS and it tended to penalise noise variables more than BHS did. When all the variables within a group had small effects, the HBGHS tended to apply extra shrinkage to the whole group and shrink them together.

We applied BHS, BGHS, HBGHS and lasso with BIC to three real datasets: a diabetes dataset, a Boston housing dataset, and an electrical-maintenance dataset, using three expansions: Legendre polynomials, Fourier expansions, Haar wavelets, and compared mean squared prediction error.

For diabetes data, the BGHS produced the smallest mean prediction error when the data is not expanded. When expanding using Legendre polynomials, BHS model with expansion of degree 3 produced the smallest prediction error. As the degree of expansions grows from degree 3, the predictive performance become worse for all models and the performance deteriorated. When expanding using Fourier expansions, the BHS model with one degree of expansion produced the smallest mean prediction error compared to all other models. The Fourier expansions do not break down as degree of expansions increased and are very stable. When expanding using Haar wavelets, although all methods with degree greater

than one did not return a smaller values of prediction error compared to BHS without expansion, their performance did not deteriorate for a relatively high degree of expansions. The BGHS did not work as well as the BHS or HBGHS when degrees of expansions were large because the BGHS do not perform individual shrinkage within each group, and therefore, it is worse when there exist a small number of strong effects in a large group.

For Boston housing data, the BGHS with Legendre expansion of degree 1 returned the smallest mean prediction error for all methods with all degrees of Legendre expansions. For all methods with degree less than 4, the mean of prediction error for BHS-NE was no better than all Bayesian-type models with expansions which implies that there may exist non-linear relationship in predictors for Boston housing data. However, when degree increased, all methods started to break down, and therefore the Legendre expansion is not very stable. For Fourier expansions of the predictor, the BGHS consistently produced the smallest prediction errors for all degrees of expansions compared to all other models. The BHS-NE gave the largest values in terms of mean squared prediction error. Therefore, there is non-linear effects in the data, which coincides with results from the Legendre expansion. For Haar wavelet expansion, the BGHS with degree 3 produced the smallest prediction error across all degrees and all methods. The BHS-NE in this case did not gave as small mean prediction error as all four methods did. The Haar wavelet method is quite stable because the mean prediction error did not break down until the degree of expansions reaches an extremely large value.

For electrical-maintenance data, there clearly existed a nonlinear pattern in the data because the mean squared prediction error decreased as the degree of expansions increased across all methods. On average, the HBGHS had the lowest MSPE, especially when the degree of polynomials grew. The BGHS had the smallest MSPE when each group had few variables ($K = 2$).

Therefore, BGHS method and the HBGHS method is capable of performing both group-wise and within group selection. They have indicated good performance in terms of the mean squared prediction error on both simulated data and real

data. The proposed methods outperform the regular BHS method when applied to nonlinear functions and additive models in simulated examples. Even when there is no underlying group structure, the proposed methods are competitive with the regular BHS method. The results of real data analysis also support the above promising performance.

In Chapter 5, we proposed a more general Bayesian global-local shrinkage hierarchical model to handle group structures that include overlapping and multilevel group structures; in particular, we extend the recent horseshoe and horseshoe+ priors for complex grouped predictor structures. After obtaining the posterior samples from the Bayesian models, we extended the decoupled shrinkage and selection method for grouped variables and apply the group non-negative garrotte to produce sparsified posterior estimates. Generalised information criteria using our proposed posterior expected estimates of the degrees of freedom of the models in the group DSS model path are then used to select the final model, and perform group-wise selection.

We applied our method to three simulated examples and the results showed that the procedure using our proposed estimate of the degrees of freedom performed well in selecting active groups, especially when these groups were sparse. Compared to the Bayesian grouped slab-and-spike (BSGS) method, our method demonstrated similar performance in terms of the overall rate of correct group identifications, while requiring substantially less computational resources. The performance of our proposed method in terms of group selection for Example 3 shows good results as both BIC* and MMLu* select the active group all the time and selected no false groups.

However, the performance of individual selection under the small-n-large-p situation is not as good as BSGS method. Although our proposed method (NNG, KL, IC) selected true predictors all the time, it tends to pick more false positives compared to BSGS method. In all scenarios, the use of small-sample information criteria to select sparse models appeared to be preferable to the original heuristic model selection criteria proposed in [14].

Since our proposed method tends to overfit inactive groups in small-n-large-p scenario, we further extended our proposed method to allow for three variations:

- estimates of model coefficients used in the information criteria to be either least squares (LS) estimates or non-negative garrotte (NNG) estimates;
- likelihood function used in information criteria varied to be either our proposed KL measurement (KL) or conventional negative log-likelihood function (L);
- penalty function in information criteria to be usual penalty function (IC) or extended penalty function designed for small-n-large-p scenario (EIC).

Next we applied those eight combinations to same three simulated examples. Simulation results of all eight combinations showed that most methods worked well when there is moderate size of active groups. The EIC-type methods did not work well for moderate number of inactive groups because they tend to overfit the inactive groups. When the model is extreme sparse, only (NNG, KL, EIC) and (LS, KL, EIC) worked well in terms of individual selection because they did not over select the inactive groups. Both of them also performed well in group selection since they selected active groups all the time and selected no false groups.

We then applied both (NNG, KL, IC) and (LS, KL, EIC) to two real datasets: a birthweight dataset and a gene-drug dataset. The prediction errors of our proposed method (NNG, KL, IC) obtained in an experiment using real data also showed an improvement over the BSGS method and the original DSS model selection criteria. These results suggest that the proposed grouped DSS procedure with information criteria is an efficient tool for selecting sparse grouped models with good predictive performance.

In conclusion, we recommend that (NNG, KL, IC) is suitable for selecting groups, and (LS, KL, EIC) is suitable for individual selection under small-n-large-p situation.

In Chapter 6, we introduced conventional likelihood ratio (LR) test and posterior likelihood ratio (PLR) test, and experimentally explored their Type I error and power for both methods. The PLR generally had a lower Type I error rate while retaining relative competitive performance in terms of the power of the method compared to LR. A high-dimensional real GWAS dataset, the Haiman-Hopper breast cancer dataset, was then introduced and used for gene-based testing. For each gene, we fit a Bayesian group model to the data by treating principal components as a group and all SNPs within the gene as another group, and computed a series of p-values for each gene.

There were 86 genes selected using proposed method, and supporting literature has been found for their association with breast cancer. We also compared the correlation of p-values among proposed method, LR, versatile gene-based association study (VEGAS) and the generalized higher criticism (GHC). The correlation between PLR and LR is the most significant among four methods while the correlation between LR and GHC is the smallest. Most methods shown moderate to strong correlation between each other except for the correlation between LR and GHC, which implies that there are a few common genes selected by different approaches.

We also developed a predictive model to Haiman-Hopper dataset. The data was divided into training dataset and testing dataset, and the model finally selected 1579 SNPs with area under the curve 0.6544. We further reduced the number of SNPs to 1000 and the area under the curve was slightly reduced to 0.6486.

7.2 Future Work

Our proposed group DSS procedure introduced in Chapter 5 is only for selection of groups at single level, and it can be extended to select groups for multilevel. For example, after obtaining posterior samples from the Bayesian grouped model, group DSS can be run separately at each group level to select groups of variables

for every single level. Therefore, we can obtain a series of grouped variables for multilevel structures.

The limitation of our current computing technology does not allow us to fit the generalized Bayesian group model introduced in Chapter 5 to high dimensional data simultaneously, so we incorporated the posterior likelihood ratio with the Bayesian model, and even in this case, the computational cost was substantial. With the development of future technology, it is also desirable if the generalized Bayesian group model could be applied to high dimensional data and conduct the analysis with a large scale. So we can fit the proposed model to all data with underlying group structures being considered, and select individuals (SNPs) or groups (genes) simultaneously provided a more powerful computing system.

The proposed Bayesian method can also be applied to other high-dimensional genetics data, such as lung cancer, prostate cancer, colorectal cancer, etc. It has the potential to identify more genetic variants associated with increased risks, gain insights and investigate the genetic basis of more complex diseases, and be beneficial to medical treatment and cancer prevention.

A Bayesian feature ranking algorithm could be developed to produce a ranking for grouped predictors. After obtaining posterior samples from our proposed Bayesian sparse grouped models from Chapter 5, a ranking method could be built to compute rankings for all sets of grouped predictors based on posterior samples and obtain a final ranking for all groups.

For example, a ranking method can be applied to all predictors or groups of predictors for each sample, and then sum up the total rankings for all samples to acquire a final ranking. If there exist overlapped groups, a ranking could be computed according to individual predictors and add rankings for all predictors that belong to the same group together to produce group ranking for one sample, and add up group rankings for all samples to obtain a final group ranking. The advantages of an approach such as this include:

- The flexibility of the method in being able to rank individual variables, or rank groups of variables.

- The Bayesian grouped ranking method enables the applications of various models such as linear regression models and logistic regression models. In addition, various alternative prior distributions can be integrated into the algorithm easily so that it is convenient to incorporate different prior knowledge to the model and increase flexibility.

- For high dimensional data, many predictors may be associated with each other. The model fits all data simultaneously which easily takes correlation between predictors into consideration. The approach enables a joint ranking for variables, so joint information or common traits are also considered.

- In addition, it computes the credible intervals of the rankings which can be used for further analysis.

- By ranking grouped variables instead of individual predictor, the dimensionality can also be reduced sufficiently.

When applying group ranking method to GWAS datasets, explanatory variables can be divided into disjoint groups, and obtain a group ranking. There are some possible ways of dividing variables which include: divide all SNPs by linkage disequilibrium (LD) blocks; or group SNPs in the same gene, and produce ranking for each gene.

Appendix A

Top 1000 SNPs selected for Haiman-Hopper dataset

TABLE A.1: A list of top 1000 SNPs selected for H2 dataset.

Table A.1 – *Continued from previous page*

kgp4699915	kgp250377	kgp1843912	kgp12065301	kgp8561074	kgp1529490	rs2114684
kgp9381927	rs2420941	kgp7219968	rs2250218	kgp1399566	kgp153261	kgp7098078
rs1649204	kgp29677299	rs3135818	kgp1536347	rs2556537	rs3135814	kgp4344526
rs1613776	kgp5113178	kgp4359919	kgp11340734	kgp30007191	rs7898010	kgp7622803
rs3135772	rs7090018	rs2912759	kgp22834115	kgp10196762	rs3135766	kgp9066305
rs2912762	kgp9493344	kgp7640870	rs2071616	kgp3719376	kgp12293985	kgp334071
kgp4137428	rs2912770	kgp11731038	kgp4853260	kgp415236	kgp4787106	rs17614209
rs2981427	rs3135730	kgp6289055	rs4752567	kgp5387030	rs2981428	kgp413010
kgp4502919	kgp1271706	kgp4501556	kgp1054829	kgp888763	exm-rs2981579	kgp9612135
kgp687485	rs1078806	kgp9094994	kgp4383822	kgp10967515	rs45631582	rs1219643
rs17102287	rs2860197	rs2420946	kgp11763324	kgp7059655	kgp8384502	kgp11177242
kgp9597836	kgp5178869	rs1863744	rs1219639	kgp11668542	kgp3759954	kgp4350805
kgp1214471	kgp6040641	kgp7751789	kgp9631270	kgp5645690	kgp25700296	kgp11142406
kgp7974579	kgp25778580	kgp1851200	kgp4703982	kgp9702713	kgp5574790	kgp3989966
kgp4711849	rs4594251	kgp1018650	exm-rs4784227	kgp86210	rs3104758	rs9708611
rs12935019	rs4784230	rs3112633	kgp2728865	kgp10176091	kgp11635579	kgp9891417
kgp11335121	rs2335	rs3104766	rs12920540	kgp1969925	rs11075551	kgp2545498
kgp11475633	rs12922061	rs3112612	kgp25807115	kgp4865662	kgp11881114	kgp11207777
kgp6372228	kgp9076291	kgp25633661	kgp10709152	kgp9745417	kgp10852418	kgp10024193
kgp16309139	kgp8814272	kgp10161709	rs3104805	kgp2187308	kgp5286968	kgp4299519
kgp1394407	kgp124620	kgp8193080	rs1420537	kgp3610834	rs2387575	kgp5900021
rs16951139	rs11075494	kgp8840943	kgp25784757	rs3095668	kgp25653948	kgp240802
kgp25806195	kgp6357901	kgp4578972	kgp1437179	kgp12017338	kgp2311323	kgp25743380
kgp25803316	kgp5384523	exm2252620	kgp7173428	rs12597716	kgp10003036	rs3095611
kgp323210	kgp9864053	kgp287792	kgp9842684	kgp9635427	rs11647305	kgp6978599
kgp25765090	kgp6782482	rs8046780	kgp25741895	rs10852413	rs4141496	kgp16416851
kgp3286820	kgp11867617	kgp3723704	kgp8360223	kgp1325013	kgp10684221	kgp7530402
kgp2406425	kgp4565327	rs9302555	kgp16306074	rs1111481	rs8051542	rs4784220
kgp1662333	kgp11834515	kgp2921837	kgp9822552	kgp6885085	rs9933638	rs9302556
kgp9727422	rs12443621	kgp6918479	kgp10280942	kgp9978795	kgp2105325	kgp8502564
kgp4328297	kgp11568845	kgp4921604	kgp12258964	rs1420533	kgp25798421	kgp5438107
kgp7751789	kgp9631270	kgp5645690	kgp25700296	rs4783780	kgp7974579	kgp25778580
kgp1851200	kgp8588818	kgp5574790	rs3803662	kgp3989966	kgp1762679	kgp4711849
kgp1891317	kgp1018650	exm-rs4784227	rs498337	kgp2493809	rs907605	kgp12115064

Continued on next page

TABLE A.1: A list of top 1000 SNPs selected for H2 dataset.

Table A.1 – *Continued from previous page*

kgp6294358	exm877581	kgp2362929	kgp12406771	kgp638353	kgp8451788	rs626738
rs34156488	rs192955527	kgp2698478	kgp10506355	kgp5527825	rs7130886	kgp1937135
kgp9442433	kgp402647	kgp10027792	kgp10203784	rs2089910	rs73410935	rs599774
kgp1400231	kgp9093738	kgp1888333	kgp8755204	rs587961	kgp626810	rs2137320
kgp1325007	rs11041540	kgp7591145	rs622173	rs6578896	kgp5886403	kgp1648065
kgp9427035	rs4980379	rs11041553	kgp5978380	kgp9669014	kgp1462410	kgp4381527
kgp2757532	rs4980384	kgp1966669	exm877731	kgp171523	rs661348	kgp4252858
kgp4343741	rs3817197	kgp9853898	rs144344717	exm-rs3817198	rs17173834	rs542605
kgp4048410	kgp11480846	rs7483477	rs4264135	kgp12393956	kgp3576326	kgp3827191
kgp7382593	kgp28566155	kgp10673542	rs283709	kgp8214692	rs924992	kgp3591250
rs185852	rs412835	rs446123	kgp11897167	rs283701	kgp1486148	kgp7434376
kgp12118473	rs10100445	kgp11239765	rs6984900	kgp5847405	kgp766315	rs283718
kgp2750922	rs283720	rs283721	kgp9529625	rs10111296	kgp7892201	kgp2017871
kgp10083454	kgp9566551	kgp4481168	kgp10986207	rs148267414	kgp3902649	kgp10345284
rs7017081	kgp8004757	exm-rs16902094	kgp4863583	kgp760058	rs1995613	kgp28912101
kgp733976	kgp28972957	rs620861	kgp5591927	kgp7771502	kgp8916961	rs424281
rs16902104	kgp9363388	rs587948	kgp7626693	rs10956359	rs17464492	kgp7216574
kgp7112868	kgp8000269	rs672888	kgp28742733	rs391640	kgp3745402	kgp6841858
kgp9071392	kgp6687218	kgp3647819	rs12541832	kgp1861642	kgp12392257	rs13271658
rs10095860	rs10098985	rs6982616	exm-rs13281615	kgp8991091	kgp928954	rs13267780
kgp12264007	rs12156034	kgp4858858	kgp5928307	kgp2920629	kgp11451096	kgp11514851
kgp9426564	kgp10021568	kgp11222964	kgp11607992	kgp28947518	kgp1219264	rs2060775
kgp3145103	rs11776260	kgp5343774	rs9643220	exm-rs1562430	kgp807254	kgp6492077
kgp1036224	rs731900	rs16902131	kgp7346978	kgp8977984	kgp5105080	kgp984461
kgp4040823	kgp1207376	kgp4101311	kgp2457355	kgp3988448	kgp8105932	rs6986543
rs79122086	kgp7788869	kgp7604416	kgp521632	kgp6199020	rs7820981	kgp7522859
rs1562871	kgp3772919	kgp6536926	rs10505477	kgp9693043	rs10505476	rs10808555
kgp8607138	rs72550838	rs10808556	exm-rs6983267	kgp11478084	kgp12094946	rs13248944
kgp1963147	exm-rs7014346	kgp7678949	kgp9237867	rs4871022	kgp20142786	kgp12497727
kgp6624495	rs6985419	kgp10822318	rs7842552	kgp4792276	kgp777571	rs12375310
kgp7690666	kgp2210730	kgp9623506	kgp4888039	rs1447294	rs75542580	rs9297756
rs6995633	rs7000448	rs4871790	kgp1367278	kgp1584215	kgp3444125	rs6981397
kgp1025286	kgp2338492	kgp11583380	kgp10623397	kgp3383436	kgp9067147	kgp1050139
kgp7194081	kgp3357094	kgp28605828	kgp7244628	kgp11542021	kgp4044433	kgp11126791
kgp10470725	kgp6692073	rs7841264	kgp2670155	kgp7913130	kgp9590193	kgp4433843
rs10956372	rs7830412	rs10094871	kgp7485047	rs1447293	rs921146	kgp4533885
kgp9782173	kgp9841042	kgp10843896	kgp4246960	rs6981424	kgp6421341	kgp5612150
rs9297758	kgp5998564	rs10086608	kgp1774591	exm2270923	kgp10756259	kgp8508677
kgp12207774	kgp1919810	kgp22813487	kgp28577613	kgp753163	rs9643225	rs11775749
rs9643227	rs16902169	rs13258812	rs723555	kgp904739	kgp28957693	kgp2145978
rs16902173	rs12155672	rs1562432	rs1562431	rs72714552	kgp6133558	kgp8475119
kgp28854179	kgp6315408	kgp5766752	kgp4723535	kgp9638530	kgp1258573	rs4871808
kgp1407196	kgp641779	kgp177511	kgp8247117	kgp4381527	rs2271439	rs4980384
GA027546	kgp171523	rs661348	kgp4252858	kgp1756859	kgp7229739	exm-rs3817198
rs17173834	rs542605	kgp11480846	kgp4651445	rs7483477	rs4264135	rs7115847
kgp3576326	kgp1625006	kgp22837749	kgp4437273	exm-rs909116	rs2668870	kgp9850532
kgp10236367	rs1398256	rs2334385	rs965912	kgp129483	rs3781957	rs3781956
kgp2750320	kgp3916820	kgp3917446	kgp10719519	kgp11443471	kgp2427943	kgp8442735
kgp6723579	kgp7821836	rs13259126	kgp3177285	kgp1490057	kgp3059858	kgp760721
rs17377068	kgp3046097	rs283741	kgp9717727	kgp11139791	kgp10028406	kgp3499017
rs283735	rs4871014	kgp7875824	kgp4992218	kgp1096537	rs34383040	rs13281552
rs73705767	kgp2851769	kgp2279508	rs11777807	kgp10582626	rs12549405	kgp8716561
kgp8203094	rs1011387	kgp3827191	kgp7382593	kgp10673542	rs283710	kgp409844
kgp8214692	rs924992	kgp28921043	rs185852	rs412835	rs446123	kgp11897167
rs283701	kgp1486148	kgp7434376	kgp12118473	rs10100445	kgp11239765	rs6984900
rs924993	kgp4386994	rs283717	kgp2556739	kgp766315	rs11994592	rs283718
kgp2750922	kgp11689270	rs283720	rs6996866	kgp9529625	rs10111296	kgp2411054
kgp7892201	kgp2017871	kgp9566551	kgp4073596	kgp10986207	rs148267414	kgp3902649
kgp10345284	kgp8004757	exm-rs16902094	kgp760058	kgp9063092	kgp3396960	rs445114
kgp1028177	kgp733976	rs620861	kgp7771502	kgp8916961	kgp2967928	rs424281
rs16902104	kgp9363388	rs587948	kgp7626693	rs17464492	kgp7216574	kgp659122

Continued on next page

TABLE A.1: A list of top 1000 SNPs selected for H2 dataset.

Table A.1 – *Continued from previous page*

kgp7112868	rs687279	kgp3745402	rs77209413	kgp6841858	kgp9071392	kgp6687218
kgp3647819	kgp7199586	kgp1861642	kgp12392257	rs13271658	rs10095860	rs10098985
exm-rs13281615	rs13256275	kgp8991091	kgp1185388	kgp28746100	kgp12264007	kgp7355449
kgp3040784	kgp126751	kgp5928307	kgp2920629	kgp11451096	kgp10093087	kgp11514851
kgp2231562	kgp10021568	kgp11607992	kgp1219264	kgp4195808	kgp1365181	rs16902126
rs11776260	rs9643220	exm-rs1562430	kgp10956083	kgp7485203	kgp807254	kgp6492077
rs731900	kgp7346978	kgp984461	kgp4040823	kgp4101311	kgp2457355	kgp3988448
kgp4716079	kgp8105932	rs6986543	rs7820981	rs1562871	rs12549845	kgp3772919
rs7844673	kgp4837117	kgp6536926	kgp8517836	kgp11427803	rs10505477	kgp7779932
kgp676769	kgp9693043	rs10505476	rs10808555	kgp8607138	kgp4426197	kgp7957538
rs10505475	rs72550838	rs10808556	exm-rs6983267	kgp11478084	kgp1826686	rs7837328
kgp1963147	exm-rs7014346	kgp8864742	rs3999187	kgp6534891	kgp608836	rs10995132
kgp8119283	rs7080900	rs7096931	kgp2803408	kgp8347171	kgp7190227	kgp5562572
rs10733773	rs10761628	kgp2907110	rs3852425	kgp3306028	rs146182625	rs7920612
rs10740076	kgp5142792	rs2306970	exm827620	kgp1884699	kgp10989713	rs10821993
kgp5033443	rs10995147	kgp1882964	kgp5012733	kgp11841932	kgp12507633	kgp1216836
kgp8933004	rs4745969	kgp11830150	rs7912735	kgp10489122	rs4746853	rs10995153
rs4636539	kgp7877024	rs10995154	kgp6661821	kgp1576295	rs7907542	kgp1226393
kgp11391424	rs11813357	rs17312431	kgp29820871	kgp5879081	kgp8196880	kgp5141600
rs17312550	kgp29674819	kgp2947314	kgp29955256	kgp6268942	rs7074870	rs2393881
rs17221319	rs10822007	kgp8692631	rs3847337	kgp10078376	rs147765741	kgp10644911
kgp10058361	kgp3814311	rs10822008	kgp9113853	kgp931687	kgp8350274	rs149861829
rs12220488	rs11818767	kgp9111615	kgp5916227	kgp4363710	kgp10843192	kgp1435796
kgp12440802	kgp10485084	rs1955332	kgp8867568	kgp4817729	kgp1113165	kgp29721685
rs9804184	rs11819488	kgp1855958	kgp5661568	rs4282885	kgp849511	kgp1688795
rs6479818	kgp8181840	kgp10919864	rs75984279	kgp10722525	rs6479821	kgp3739845
kgp226398	kgp4500395	rs10509168	rs2393886	rs11812404	kgp2496702	kgp6427040
rs11713419	kgp10352686	rs168025	kgp12493637	kgp8677265	rs3806681	kgp10838102
kgp9527308	rs17878483	kgp11547771	kgp4985075	kgp2714239	kgp9344972	kgp3244080
rs334792	rs9814648	kgp3621078	rs7618222	kgp4265018	exm2255974	kgp6358045
kgp10767328	rs9808896	kgp7757119	kgp10320060	kgp6881245	kgp10781956	kgp9100106
kgp1411163	kgp10289280	kgp10987834	kgp22759732	kgp17766417	kgp9197516	kgp5834303
kgp3649291	rs1278150	kgp4999947	kgp5561650	rs184087127	kgp573043	kgp9555259
kgp26622554	kgp3988792	kgp10227281	rs2630819	kgp1901081	kgp4229150	kgp1843741
rs7638187	kgp5645214	kgp10317445	rs334763	kgp5466401	kgp1069595	kgp26743336
rs3804783	kgp10744991	kgp414387	kgp4191859	kgp27055230	rs1497	kgp1502022
kgp4674185	rs1672770	rs1620675	kgp10900846	kgp4125872	kgp26625832	rs1672767
rs1669338	rs1705832	rs1669336	kgp6989138	kgp3793157	rs17027750	kgp997739
rs711619	kgp3736637	kgp3445495	rs4685611	rs3792414	kgp7171662	kgp26664310
rs10745367	rs11103084	rs180984113	kgp5246821	kgp6042990	kgp2941321	rs872005
kgp5204052	kgp10405704	rs11792145	kgp371894	kgp5285295	kgp971037	kgp11877500
kgp3847935	kgp8136657	kgp8576193	kgp1154347	rs4841889	kgp5096829	kgp9897767
rs11103105	kgp8706762	rs12336965	kgp10574259	exm796073	exm-IND9-137656167	rs12347803
rs11103106	rs10858139	kgp5645661	kgp2027989	rs7870610	rs1333235	rs1333234
kgp2255654	kgp9424088	kgp10348906	kgp11590035	kgp9888611	kgp3583242	rs11103115
kgp10506793	kgp1865548	rs7855164	kgp12552654	kgp10171893	rs13292721	rs12684457
kgp6167443	rs13286163	rs11103123	kgp2822305	kgp11404093	kgp9460833	kgp5614171
kgp9756240	kgp10708457	rs4842019	kgp11496967	kgp27384948	kgp7269910	

Bibliography

[1] K. McPherson, C. Steel, and J. M. Dixon, "ABC of breast diseases: Breast cancer - epidemiology, risk factors, and genetics," *BMJ: British Medical Journal*, vol. 321, no. 7261, pp. 624–628, 2000.

[2] E. M. L. Beale, M. G. Kendall, and D. W. Mann, "The discarding of variables in multivariate analysis," *Biometrika*, vol. 54, no. 3/4, pp. 357–366, 1967.

[3] R. R. Hocking and R. N. Leslie, "Selection of the best subset in regression analysis," *Technometrics*, vol. 9, no. 4, pp. 531–540, 1967.

[4] N. R. Draper and H. Smith, *Applied regression analysis*. New York: John Wiley and Sons, 1966.

[5] J. M. Sutter and J. H. Kalivas, "Comparison of forward selection, backward elimination, and generalized simulated annealing for variable selection," *Microchemical Journal*, vol. 47, no. 1-2, pp. 60–66, 1993.

[6] C. Leng, Y. Ling, and G. Wahba, "A note on the Lasso and related procedures in model selection," *Statistica Sinica*, vol. 16, no. 4, pp. 1273–1284, 2006.

[7] L. Breiman, "Better subset regression using the nonnegative garrote," *Technometrics*, vol. 37, no. 4, pp. 373–384, 1995.

[8] A. E. Hoerl and R. W. Kennard, "Ridge regression: Biased estimation for nonorthogonal problems," *Technometrics*, vol. 12, no. 1, pp. 55–67, 1970.

[9] R. Tibshirani, "Regression shrinkage and selection via the Lasso," *Journal of the Royal Statistical Soiety. Series B (Methodological)*, vol. 58, no. 1, pp. 267–288, 1996.

[10] J. Friedman, T. Hastie, and R. Tibshirani, "A note on the group Lasso and a sparse group Lasso," *arXiv preprint arXiv:1001.0736*, 2010.

[11] H. Zou, "The adaptive Lasso and its oracle properties," *Journal of the American statistical association*, vol. 101, no. 476, pp. 1418–1429, 2006.

[12] N. G. Polson and J. G. Scott, "Shrink globally, act locally: sparse Bayesian regularization and prediction," *Bayesian Statistics*, vol. 9, pp. 501–538, 2010.

[13] A. Bhattacharya, D. Pati, N. S. Pillai, and D. B. Dunson, "Dirichlet–Laplace priors for optimal shrinkage," *Journal of the American Statistical Association*, vol. 110, no. 512, pp. 1479–1490, 2015.

[14] P. R. Hahn and C. M. Carvalho, "Decoupling shrinkage and selection in Bayesian linear models: a posterior summary perspective," *Journal of the American Statistical Association*, vol. 110, no. 509, pp. 435–448, 2015.

[15] M. Bayes and M. Price, "An essay towards solving a problem in the doctrine of chances. By the late Rev. Mr. Bayes, F. R. S. communicated by Mr. Price, in a letter to John Canton, A. M. F. R. S.," *Philosophical Transactions (1683-1775)*, vol. 53, pp. 370–418, 1763.

[16] H. Jeffreys, *Theory of probability*. Oxford University Press, 3 ed., 1961.

[17] J. M. Bernardo, "Reference posterior distributions for Bayesian inference," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 41, no. 2, pp. 113–128, 1979.

[18] S. Kullback and R. A. Leibler, "On information and sufficiency," *The Annals of Mathematical Statistics*, vol. 22, no. 1, pp. 79–86, 1951.

[19] R. E. Kass and A. E. Raftery, "Bayes factors," *Journal of the American Statistical Association*, vol. 90, no. 430, pp. 773–795, 1995.

[20] S. M. Lewis and A. E. Raftery, "Estimating Bayes estimators via posterior simulation with the Laplace-Metropolis estimator," *Journal of the American Statistical Association*, vol. 92, no. 438, pp. 648–655, 1997.

[21] M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*. Dover Publications, 1972.

[22] P. C. Hammer, "The midpoint method of numerical integration," *Mathematics Magazine*, vol. 31, no. 4, pp. 193–195, 1958.

[23] C. M. Bishop *et al.*, *Neural networks for pattern recognition*. Oxford University Press, 1995.

[24] D. J. MacKay, *Information theory, inference and learning algorithms*. Cambridge University Press, 2003.

[25] A. E. Gelfand and A. F. M. Smith, "Sampling-based approaches to calculating marginal densities," *Journal of the American statistical association*, vol. 85, no. 410, pp. 398–409, 1990.

[26] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, "Equation of state calculations by fast computing machines," *The Journal of Chemical Physics*, vol. 21, no. 6, pp. 1087–1092, 1953.

[27] W. K. Hastings, "Monte carlo sampling methods using markov chains and their applications," *Biometrika*, vol. 57, pp. 97–109, 1970.

[28] S. Geman and D. Geman, "Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 6, no. 6, pp. 721–741, 1984.

[29] P. McCullagh and J. A. Nelder, *Generalized Linear Models, Second Edition*. Chapman and Hall/CRC, 1989.

[30] B. K. Natarajan, "Sparse approximate solutions to linear systems," *SIAM Journal on Computing*, vol. 24, no. 2, pp. 227–234, 1995.

[31] D. Bertsimas, A. King, and R. Mazumder, "Best subset selection via a modern optimization lens," *The Annals of Statistics*, vol. 44, no. 2, pp. 813–852, 2016.

[32] T. Hastie, R. Tibshirani, and R. J. Tibshirani, "Extended comparisons of best subset selection, forward stepwise selection, and the Lasso," *arXiv preprint arXiv:1707.08692*, 2017.

[33] C. Stein, "Inadmissibility of the usual estimator for the mean of a multivariate normal distribution," *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1, pp. 197–206, 1956.

[34] W. James and C. Stein, "Estimation with quadratic loss," *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, vol. 1, pp. 361–379, 1961.

[35] S. L. Sclove, "Improved estimators for coefficients in linear regression," *Journal of the American Statistical Association*, vol. 63, no. 322, pp. 596–606, 1968.

[36] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 67, no. 2, pp. 301–320, 2005.

[37] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 68, no. 1, pp. 49–67, 2006.

[38] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, "Least angle regression," *The Annals of Statistics*, vol. 32, no. 2, pp. 407–499, 2004.

[39] J. Friedman, T. Hastie, and T. Tibshirani, "Regularization paths for generalized linear models via coordinate descent," *Journal of Statistical Software*, vol. 33, no. 1, pp. 1–22, 2010.

[40] J. Friedman, T. Hastie, H. Hofling, and R. Tibshirani, "Pathwise coordinate optimization," *The Annals of Applied Statistics*, vol. 1, no. 2, pp. 302–332, 2007.

[41] D. Gabay and B. Mercier, "A dual algorithm for the solution of nonlinear variational problems via finite element approximation," *Computers & Mathematics with Applications*, vol. 2, no. 1, pp. 17–40, 1976.

[42] S. Boyd, N. Parikh, E. Chu, B. Peleato, J. Eckstein, *et al.*, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Foundations and Trends® in Machine learning*, vol. 3, no. 1, pp. 1–122, 2011.

[43] M. R. Hestenes, "Multiplier and gradient methods," *Journal of Optimization Theory and Applications*, vol. 4, no. 5, pp. 303–320, 1969.

[44] H. Everett III, "Generalized Lagrange multiplier method for solving problems of optimum allocation of resources," *Operations Research*, vol. 11, no. 3, pp. 399–417, 1963.

[45] J. Fan and R. Li, "Variable selection via nonconcave penalized likelihood and its oracle properties," *Journal of the American statistical Association*, vol. 96, no. 456, pp. 1348–1360, 2001.

[46] H. Zou and R. Li, "One-step sparse estimates in nonconcave penalized likelihood models," *The Annals of Statistics*, vol. 36, no. 4, pp. 1509–1533, 2008.

[47] H. Akaike, "Information theory and an extension of the maximum likelihood principle," *Second International Symposium on Information Theory*, vol. I, pp. 267–281, 1973.

[48] G. Schwarz, "Estimating the dimension of a model," *The Annals of Statistics*, vol. 6, no. 2, pp. 464–464, 1978.

[49] J. E. Cavanaugh, "A large-sample model selection criterion based on kullback's symmetric divergence," *Statistics & Probability Letters*, vol. 42, no. 4, pp. 333–343, 1999.

[50] C. S. Wallace and D. M. Boulton, "An information measure for classification," *The Computer Journal*, vol. 11, no. 2, pp. 185–194, 1968.

[51] C. M. Hurvich and C.-L. Tsai, "Regression and time series model selection in small samples," *Biometrika*, vol. 76, no. 2, pp. 297–307, 1989.

[52] E. Makalic and D. F. Schmidt, "Efficient linear regression by minimum message length," tech. rep., Monash University, Tech. Rep, 2006.

[53] D. F. Schmidt and E. Makalic, "Mml invariant linear regression," in *Australasian Joint Conference on Artificial Intelligence*, pp. 312–321, 2009.

[54] D. V. Lindley and A. F. M. Smith, "Bayes estimates for the linear model," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 34, no. 1, pp. 1–41, 1972.

[55] T. J. Mitchell and J. J. Beauchamp, "Bayesian variable selection in linear regression," *Journal of the American Statistical Association*, vol. 83, no. 404, pp. 1023–1032, 1988.

[56] E. I. George and R. E. McCulloch, "Approaches for Bayesian variable selection," *Statistica Sinica*, vol. 7, no. 2, pp. 339–373, 1997.

[57] V. E. Johnson and D. Rossell, "Bayesian model selection in high-dimensional settings," *Journal of the American Statistical Association*, vol. 107, no. 498, pp. 649–660, 2012.

[58] N. G. Polson, J. G. Scott, and J. Windle, "The bayesian bridge," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 76, no. 4, pp. 713–733, 2014.

[59] T. Park and G. Casella, "The Bayesian Lasso," *Journal of the American Statistical Association*, vol. 103, no. 482, pp. 681–686, 2008.

[60] C. Leng, M.-N. Tran, and D. Nott, "Bayesian adaptive Lasso," *Annals of the Institute of Statistical Mathematics*, vol. 66, no. 2, pp. 221–244, 2014.

[61] Q. Li, N. Lin, *et al.*, "The bayesian elastic net," *Bayesian analysis*, vol. 5, no. 1, pp. 151–170, 2010.

[62] C. M. Carvalho, N. G. Polson, and J. G. Scott, "The horseshoe estimator for sparse signals," *Biometrika*, vol. 97, no. 2, pp. 465–480, 2010.

[63] A. Bhadra, J. Datta, N. G. Polson, and B. Willard, "The horseshoe+ estimator of ultra-sparse signals," *Bayesian Analysis*, vol. 12, no. 4, pp. 1105–1131, 2017.

[64] C. Hans, "Bayesian lasso regression," *Biometrika*, vol. 96, no. 4, pp. 835–845, 2009.

[65] C. M. Carvalho, N. G. Polson, and J. G. Scott, "Handling sparsity via the horseshoe," in *Proceedings of Machine Learning Research* (D. van Dyk and M. Welling, eds.), vol. 5, pp. 73–80, 2009.

[66] Y. Zhang, B. J. Reich, and H. D. Bondell, "High dimensional linear regression via the R2-D2 shrinkage prior," *arXiv preprint arXiv:1609.00046*, 2016.

[67] E. Makalic and D. F. Schmidt, "A simple sampler for the horseshoe estimator," *IEEE Signal Processing Letters*, vol. 23, no. 1, pp. 179–182, 2016.

[68] H. D. Bondell and B. J. Reich, "Consistent high-dimensional Bayesian variable selection via penalized credible regions," *Journal of the American Statistical Association*, vol. 107, no. 500, pp. 1610–1624, 2012.

[69] E. Makalic and D. F. Schmidt, "High-dimensional Bayesian regularised regression with the bayesreg package," *arXiv preprint arXiv:1611.06649*, 2016.

[70] X. Tang, X. Xu, M. Ghosh, and P. Ghosh, "Bayesian variable selection and estimation based on global-local shrinkage priors," *Sankhya A*, vol. 80, no. 2, pp. 215–246, 2018.

[71] M. Yuan and Y. Lin, "On the non-negative garrotte estimator," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 69, no. 2, pp. 143–161, 2007.

[72] H. Wang and C. Leng, "A note on adaptive group Lasso," *Computational Statistics & Data Analysis*, vol. 52, no. 12, pp. 5277–5286, 2008.

[73] F. Wei and H. Jian, "Consistent group selection in high-dimensional linear regression," *Bernoulli: Official Journal of the Bernoulli Society for Mathematical Statistics and Probability*, vol. 16, no. 4, pp. 1369–1384, 2010.

[74] S. Raman, T. J. Fuchs, P. J. Wild, E. Dahl, and V. Roth, "The Bayesian group-Lasso for analyzing contingency tables," *Proceedings of the 26th Annual International Conference on Machine Learning*, pp. 881–888, 2009.

[75] M. Kyung, J. Gill, M. Ghosh, and G. Casella, "Penalized regression, standard errors, and bayesian Lassos," *Bayesian Analysis*, vol. 5, no. 2, pp. 369–411, 2010.

[76] X. Xu and M. Ghosh, "Bayeisan variableb selection and estimation for group Lasso," *Bayesian Analysis*, vol. 10, no. 4, pp. 909–936, 2015.

[77] R.-B. Chen, C.-H. Chu, S. Yuan, and Y. N. Wu, "Bayesian sparse group selection," *Journal of Computational and Graphical Statistics*, vol. 25, no. 3, pp. 665–683, 2016.

[78] N. M. Laird and C. Lange, "Introduction to statistical genetics and background in molecular genetics," in *The Fundamentals of Modern Statistical Genetics*, pp. 1–13, Springer, 2011.

[79] N. J. Risch, "Searching for genetic determinants in the new millennium," *Nature*, vol. 405, no. 6788, pp. 847–856, 2000.

[80] A. D. Roses, "Pharmacogenetics and the practice of medicine," *Nature*, vol. 405, no. 6788, pp. 857–865, 2000.

[81] J. Drews and S. Ryser, "The role of innovation in drug development," *Nature Biotechnology*, vol. 15, no. 13, pp. 1318–1319, 1997.

[82] M. McCarthy, "Susceptibility gene discovery for common metabolic and endocrine traits," *Journal of Molecular Endocrinology*, vol. 28, no. 1, pp. 1–17, 2002.

[83] L. J. Palmer and W. O. Cookson, "Genomic approaches to understanding asthma," *Genome Research*, vol. 10, no. 9, pp. 1280–1287, 2000.

[84] L. Rodriguez-Murillo and D. A. Greenberg, "Genetic association analysis: a primer on how it works, its strengths and its weaknesses," *International journal of andrology*, vol. 31, no. 6, pp. 546–556, 2008.

[85] L. R. Cardon and L. J. Palmer, "Population stratification and spurious allelic association," *The Lancet*, vol. 361, no. 9357, pp. 598–604, 2003.

[86] E. K. Silverman and L. J. Palmer, "Case-control association studies for the genetics of complex respiratory diseases," *American Journal of Respiratory Cell and Molecular Biology*, vol. 22, no. 6, pp. 645–648, 2000.

[87] N. Risch and J. Teng, "The relative power of family-based and case-control designs for linkage disequilibrium studies of complex human diseases i. dna pooling," *Genome Research*, vol. 8, no. 12, pp. 1273–1288, 1998.

[88] A. L. Price, N. A. Zaitlen, D. Reich, and N. Patterson, "New approaches to population stratification in genome-wide association studies," *Nature Reviews Genetics*, vol. 11, no. 7, pp. 459–463, 2010.

[89] A. L. Price, N. J. Patterson, R. M. Plenge, M. E. Weinblatt, N. A. Shadick, and D. Reich, "Principal components analysis corrects for stratification in genome-wide association studies," *Nature Genetics*, vol. 38, no. 8, pp. 904–909, 2006.

[90] F.-X. Du, A. C. Clutter, and M. M. Lohuis, "Characterizing linkage disequilibrium in pig populations," *International Journal of Biological Sciences*, vol. 3, no. 3, p. 166, 2007.

[91] I. G. A. Elisseeff, "An introduction to variable and feature selection," *The Journal of Machine Learning Research*, vol. 3, pp. 1157–1182, 2003.

[92] K. Pearson, "X. on the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling," *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, vol. 50, no. 302, pp. 157–175, 1900.

[93] R. A. Fisher, "On the interpretation of χ 2 from contingency tables, and the calculation of p," *Journal of the Royal Statistical Society*, vol. 85, no. 1, pp. 87–94, 1922.

[94] Y. Benjamini and Y. Hochberg, "Controlling the false discovery rate: a practical and powerful approach to multiple testing," *Journal of the Royal Statistical Soiety*, vol. 57, no. 1, pp. 289–300, 1995.

[95] C. Li, "A mini review for aggregating analyses for genome-wide association studies (gwas)," *Biometrics & Biostatistics International Journal*, vol. 2, no. 4, pp. 115–116, 2015.

[96] P. Donnelly, "Progress and challenges in genome-wide association studies in humans," *Nature*, vol. 456, no. 7223, pp. 728–731, 2008.

[97] C. J. Hoggart, J. C. Whittaker, M. D. Iorio, and D. J. Balding, "Simultaneous analysis of all snps in genome-wide and re-sequencing association studies," *PLoS Genetics*, vol. 4, no. 7, p. e1000130, 2008.

[98] J. Li, K. Das, G. Fu, R. Li, and R. Wu, "The Bayesian Lasso for genome-wide association studies," *Bioinformatics*, vol. 27, no. 4, pp. 516–523, 2011.

[99] lldiko E. Franka and J. H. Friedman, "A statistical view of some chemometrics regression tools," *Technometrics*, vol. 35, no. 2, pp. 109–135, 1993.

[100] T. T. Wu, Y. F. Chen, T. Hastie, E. Sobel, and K. Lange, "Genome-wide association analysis by Lasso penalized logistic regression," *Bioinformatics*, vol. 25, no. 6, pp. 714–721, 2009.

[101] B. A. Logsdon, G. E. Hoffman, and J. G. Mezey, "A variational bayes algorithm for fast and accurate multiple locus genome-wide association analysis," *BMC bioinformatics*, vol. 11, no. 58, pp. 1–13, 2010.

[102] J. Fan and J. Lv, "Sure independence screening for ultrahigh dimensional feature space," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 70, no. 5, pp. 849–911, 2008.

[103] D. Paul, E. Bair, T. Hastie, and R. Tibshirani, ""Preconditioning" for feature selection and regression in high-dimensional problems," *The Annals of Statistics*, pp. 1595–1618, 2008.

[104] T. Yang, P. Thompson, S. Zhao, and J. Ye, "Identifying genetic risk factors via sparse group Lasso with group graph structure," *arXiv preprint arXiv:1709.03645*, 2017.

[105] J. Ferlay, I. Soerjomataram, R. Dikshit, S. Eser, C. Mathers, M. Rebelo, D. M. Parkin, D. Forman, and F. Bray, "Cancer incidence and mortality worldwide: sources, methods and major patterns in globocan 2012," *International Journal of Cancer*, vol. 136, no. 5, pp. E359–386, 2015.

[106] E. A. Rakha and S. Chan, "Metastatic triple-negative breast cancer," *Clinical Oncology*, vol. 23, no. 9, pp. 587–600, 2011.

[107] W. F. Anderson, N. Chatterjee, W. B. Ershler, and O. W. Brawley, "Estrogen receptor breast cancer phenotypes in the surveillance, epidemiology, and end results database," *Breast Cancer Research and Treatment*, vol. 76, no. 1, pp. 27–36, 2002.

[108] C. M. Perou *et al.*, "Molecular portraits of human breast tumours," *Nature*, vol. 406, no. 6797, p. 747, 2000.

[109] A. C. Wolff *et al.*, "American society of clinical oncology/college of american pathologists guideline recommendations for human epidermal growth factor receptor 2 testing in breast cancer," *Archives of Pathology & Laboratory Medicine*, vol. 131, no. 1, pp. 18–43, 2007.

[110] T. Sørlie *et al.*, "Repeated observation of breast tumor subtypes in independent gene expression data sets," *Proceedings of the national academy of sciences*, vol. 100, no. 14, pp. 8418–8423, 2003.

[111] P. D. P. Pharoah, N. E. Day, S. Duffy, D. F. Easton, and B. A. J. Ponder, "Family history and the risk of breast cancer: a systematic review and meta-analysis," *International Journal of Cancer*, vol. 71, no. 5, pp. 800–809, 1997.

[112] T. Walsh *et al.*, "Spectrum of mutations in brca1, brca2, chek2, and tp53 in families at high risk of breast cancer," *JAMA*, vol. 295, no. 12, pp. 1379–1388, 2006.

[113] Y. Miki, J. Swensen, D. Shattuck-Eidens, P. A. Futreal, *et al.*, "A strong candidate for the breast and ovarian cancer susceptibility gene BRCA1," *Science*, vol. 266, no. 5182, pp. 66–71, 1994.

[114] R. Wooster, G. Bignelland, J. Lancaster, S. Swiftand, *et al.*, "Identification of the breast cancer susceptibility gene BRCA2," *Nature*, vol. 378, no. 6559, pp. 789–792, 1995.

[115] C. Capalbo *et al.*, "Brca1 and brca2 genetic testing in italian breast and/or ovarian cancer families: mutation spectrum and prevalence and analysis of mutation prediction models," *Annals of Oncology*, vol. 17, no. suppl_7, pp. vii34–vii40, 2006.

[116] P. Economopoulou, G. Dimitriadis, and A. Psyrri, "Beyond brca: new hereditary breast cancer susceptibility genes," *Cancer Treatment Reviews*, vol. 41, no. 1, pp. 1–8, 2015.

[117] P. Apostolou and F. Fostira, "Hereditary breast cancer: the era of new susceptibility genes," *BioMed Research International*, 2013.

[118] J. Li *et al.*, "Pten, a putative protein tyrosine phosphatase gene mutated in human brain, breast, and prostate cancer," *Science*, vol. 275, no. 5308, pp. 1943–1947, 1997.

[119] C. Eng, "Role of pten, a lipid phosphatase upstream effector of protein kinase b, in epithelial thyroid carcinogenesis," *Annals of the New York Academy of Sciences*, vol. 968, no. 1, pp. 213–221, 2002.

[120] M. H. Gail, L. A. Brinton, D. P. Byar, D. K. Corle, S. B. Green, C. Schairer, and J. J. Mulvihill, "Projecting individualized probabilities of developing breast cancer for white females who are being examined annually," *JNCI: Journal of the National Cancer Institute*, vol. 81, no. 24, pp. 1879–1886, 1989.

[121] D. Easton *et al.*, "Genome-wide association study identifies novel breast cancer susceptibility loci," *Nature*, vol. 447, no. 7148, pp. 1087–1093, 2007.

[122] D. Hunter *et al.*, "A genome-wide association study identifies alleles in FGFR2 associated with risk of sporadic post-menopausal breast cancer," *Nature Genetics*, vol. 39, no. 7, pp. 870–874, 2007.

[123] S. N. Stacey *et al.*, "Common variants on chromosomes 2q35 and 16q12 confer susceptibility to estrogen receptor-positive breast cancer.," *Nature Genetics*, vol. 39, no. 7, pp. 865–869, 2007.

[124] S. N. Stacey, "Common variants on chromosome 5p12 confer susceptibility to estrogen receptor-positive breast cancer.," *Nature Genetics*, vol. 40, no. 6, pp. 703–706, 2008.

[125] B. Gold *et al.*, "Genome-wide association study provides evidence for a breast cancer risk locus at 6q22.33," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 105, no. 11, pp. 4340–4345, 2008.

[126] S. Ahmed, "Newly discovered breast cancer susceptibility loci on 3p24 and 17q23.2," *Nature Genetics*, vol. 41, no. 5, pp. 585–590, 2009.

[127] G. Thomas *et al.*, "A multistage genome-wide association study in breast cancer identifies two new risk alleles at 1p11.2 and 14q24.1 (rad51l1)," *Nature Genetics*, vol. 41, no. 5, pp. 579–584, 2009.

[128] W. Zhang *et al.*, "Genome-wide association study identifies a new breast cancer susceptibility locus at 6q25.1," *Nature Genetics*, vol. 41, no. 3, pp. 324–328, 2009.

[129] C. Turnbull, S. Ahmed, J. Morrison, D. Pernet, A. Renwick, M. Maranian, S. Seal, M. Ghoussaini, S. Hines, C. S. Healey, D. Hughes, M. Warren-Perry, W. Tapper, D. Eccles, D. G. Evans, *et al.*, "Genome-wide association study identifies five new breast cancer susceptibility loci," *Nature Genetics*, vol. 42, no. 6, pp. 504–507, 2010.

[130] A. C. Antoniou *et al.*, "A locus on 19p13 modifies risk of breast cancer in brca1 mutation carriers and is associated with hormone receptor-negative breast cancer in the general population," *Nature Genetics*, vol. 42, no. 10, pp. 885–892, 2010.

[131] C. A. Haiman *et al.*, "A common variant at the TERT-CLPTM1L locus is associated with estrogen receptor-negative breast cancer," *Nature Genetics*, vol. 43, no. 12, pp. 1210–1214, 2011.

[132] Q. Cai, "Genome-wide association study identifies breast cancer risk variant at 10q21.2: results from the Asia Breast Cancer Consortium," *Human Molecular Genetics*, vol. 20, no. 24, pp. 4991–4999, 2011.

[133] O. Fletcher *et al.*, "Novel breast cancer susceptibility locus at 9q31.2: results of a genome-wide association study," *Journal of the National Cancer Institute*, vol. 103, no. 5, pp. 425–435, 2011.

[134] J. Long *et al.*, "Genome-wide association study in East Asians identifies novel susceptibility loci for breast cancer," *PLoS Genetics*, vol. 8, no. 2, p. e1002532, 2012.

[135] H.-C. Kim *et al.*, "A genome-wide association study identifies a breast cancer risk variant in ERBB4 at 2q34: results from the Seoul Breast Cancer Study," *Breast Cancer Research*, vol. 14, no. 2, p. R56, 2012.

[136] A. Siddiq *et al.*, "A meta-analysis of genome-wide association studies of breast cancer identifies two novel susceptibility loci at 6q14 and 20q11," *Human Molecular Genetics*, vol. 21, no. 24, pp. 5373–5384, 2012.

[137] P. Bühlmann and S. V. D. Geer, *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media, 2011.

[138] P. Zhao, G. Rocha, and B. Yu, "The composite absolute penalties family for grouped and hierarchical variable selection," *The Annals of Statistics*, vol. 37, no. 6A, pp. 3468–3497, 2009.

[139] J. Huang, S. Ma, H. Xie, and C.-H. Zhang, "A group bridge approach for variable selection," *Biometrika*, vol. 96, no. 2, pp. 339–355, 2009.

[140] P. Breheny and J. Huang, "Penalized methods for bi-level variable selection," *Statistics and its interface*, vol. 2, no. 3, pp. 369–380, 2009.

[141] H. Dym and H. M. Jr, "Fourier series and integrals," tech. rep., 1972.

[142] R. E. Edwards, *Fourier series: a modern introduction*, vol. 2. Springer Science & Business Media, 2012.

[143] A. Haar, "Zur theorie der orthogonalen funktionensysteme," *Mathematische Annalen*, vol. 69, no. 3, pp. 331–371, 1910.

[144] J. Alcalá, A. Fernández, J. Luengo, J. Derrac, S. García, L. Sánchez, and F. Herrera, "Keel data-mining software tool: Data set repository, integration of algorithms and experimental analysis framework," *Journal of Multiple-Valued Logic and Soft Computing*, vol. 17, no. 2-3, pp. 255–287, 2010.

[145] E. I. George and R. E. McCulloch, "Variable selection via Gibbs sampling," *Journal of the American Statistical Association*, vol. 88, no. 423, pp. 881–889, 1993.

[146] M. Stone, "Cross-validatory choice and assessment of statistical predictions," *Journal of the Royal Statistical Society. Series B (Methodological)*, no. 2, pp. 111–147, 1974.

[147] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *International Joint Conference on Artificial Intelligence*, vol. 14, pp. 1137–1145, 1995.

[148] J. Shao, "Linear model selection by cross-validation," *Journal of the American statistical Association*, vol. 88, no. 422, pp. 486–494, 1993.

[149] M. J. Bayarri and J. O. Berger, "The interplay of Bayesian and frequentist analysis," *Statistical Science*, vol. 19, no. 1, pp. 58–80, 2004.

[150] Y. Zhang and H. D. Bondell, "Variable selection via penalized credible regions with Dirichlet–Laplace global-local shrinkage priors," *arXiv preprint arXiv:1602.01160*, vol. 13, no. 3, pp. 823–844, 2016.

[151] S. V. D. Pas, B. Kleijn, and A. V. D. Vaart, "The horseshoe estimator: posterior concentration around nearly black vectors," *Eectronic Journal of Statistics*, vol. 8, no. 2, pp. 2585–2618, 2014.

[152] M. Kyung, J. Gill, M. Ghosh, and G. Casella, "Penalized regression, standard errors, and Bayesian Lassos," *Bayesian Analysis*, vol. 5, no. 2, pp. 369–412, 2010.

[153] Z. Xu, D. F. Schmidt, E. Makalic, G. Qian, and J. L. Hopper, "Bayesian grouped horseshoe regression with application to additive models," in *Australasian Joint Conference on Artificial Intelligence*, pp. 229–240, Springer, 2016.

[154] C. S. Wallace, "Statistical and inductive inference by minimum message length," *Springer Science & Business Media*, 2005.

[155] J. Rissanen, "A universal prior for integers and estimation by minimum description length," *The Annals of Statistics*, vol. 11, no. 2, pp. 416–431, 1983.

[156] C. D. Giurcăneanu, S. A. Razavi, and A. Liski, "Variable selection in linear regression: Several approaches based on normalized maximum likelihood," *Signal Processing*, vol. 91, no. 8, pp. 1671–1692, 2011.

[157] A. D. R. McQuarrie and C.-L. Tsai, *Regression and time series model selection*. World Scientific, 1998.

[158] J. Friedman, T. Hastie, and R. Tibshirani, *The elements of statistical learning*. Springer series in statistics New York, 2001.

[159] J. Chen and Z. Chen, "Extended Bayesian information criterion for model selection with large model space," *Biometrika*, vol. 95, no. 3, pp. 759–771, 2008.

[160] D. W. Homser and S. Lemeshow, *Applied logistic regression*. New York: John Wiley and Sons, 1989.

[161] A. Farcomeni, "Bayesian constrained variable selection," *Statistica Sinica*, vol. 20, no. 3, pp. 1043–1062, 2010.

[162] H. Ogata, S. Goto, W. Fujibuchi, and M. Kanehisa, "Computation with the kegg pathway database," *Biosystems*, vol. 47, no. 1-2, pp. 119–128, 1998.

[163] M. Kanehisa, "The kegg database," in *'In Silico' Simulation of Biological Processes: Novartis Foundation Symposium 247*, vol. 247, pp. 91–103, Wiley Online Library, 2002.

[164] M. Kanehisa, S. Goto, S. Kawashima, Y. Okuno, and M. Hattori, "The kegg resource for deciphering the genome," *Nucleic acids research*, vol. 32, no. suppl_1, pp. 277–280, 2004.

[165] H. Jeffreys, "Some tests of significance, treated by the theory of probability," in *Mathematical Proceedings of the Cambridge Philosophical Society*, vol. 31, pp. 203–222, Cambridge University Press, 1935.

[166] A. P. Dempster, "The direct use of likelihood for significance testing," in *Proceeding of Conference on Foundational Questions in Statistical Inference*, pp. 335–352, University of Aarhus, 1974.

[167] A. P. Dempster, "The direct use of likelihood for significance testing," *Statistics and Computing*, vol. 7, no. 4, pp. 247–252, 1997.

[168] M. Aitkin, "The calibration of p-values, posterior bayes factors and the aic from the posterior distribution of the likelihood," *Statistics and Computing*, vol. 7, no. 4, pp. 253–261, 1997.

[169] X. Zhu, S. Zhang, H. Zhao, and R. S. Cooper, "Association mapping, using a mixture model for complex traits," *Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society*, vol. 23, no. 2, pp. 181–196, 2002.

[170] H. S. Chen, X. Zhu, H. Zhao, and S. Zhang, "Qualitative semi-parametric test for genetic associations in case-control designs under structured populations," *Annals of human genetics*, vol. 67, no. 3, pp. 250–264, 2003.

[171] M. N. C. Fletcher *et al.*, "Master regulators of FGFR2 signalling and breast cancer risk," *Nature Communications*, vol. 4, p. 2464, 2013.

[172] J. S. Baxter *et al.*, "Capture Hi-C identifies putative target genes at 33 breast cancer risk loci," *Nature Communications*, vol. 9, no. 1, p. 1028, 2018.

[173] J. Long *et al.*, "Identification of a functional genetic variant at 16q12. 1 for breast cancer risk: results from the Asia Breast Cancer Consortium," *PLoS Genetics*, vol. 6, no. 6, p. e1001002, 2010.

[174] Y.-J. Han, J. Zhang, Y. Zheng, D. Huo, and O. I. Olopade, "Genetic and epigenetic regulation of TOX3 expression in breast cancer," *PloS one*, vol. 11, no. 11, p. e0165559, 2016.

[175] L. Zhang and X. Long, "Association of three SNPs in TOX3 and breast cancer risk: Evidence from 97275 cases and 128686 controls," *Scientific Reports*, vol. 5, p. 12773, 2015.

[176] Y. Xu, Q. Yuan, J. Zhou, X. Chang, K. Wang, and J. Han, "Association of TOX3 polymorphisms with breast cancer: A meta-analysis," *Meta Gene*, vol. 13, pp. 70–77, 2017.

[177] H. Chen, X. Qi, P. Qiu, and J. Zhao, "Correlation between LSP1 polymorphisms and the susceptibility to breast cancer," *International Journal of Clinical and Experimental Pathology*, vol. 8, no. 5, pp. 5798–5802, 2015.

[178] J. Tang, H. Li, J. Luo, H. Mei, L. Peng, and X. Li, "The LSP1 rs3817198 T¿C polymorphism contributes to increased breast cancer risk: A meta-analysis of twelve studies," *Oncotarget*, vol. 7, no. 39, pp. 63960–63967, 2016.

[179] Y. Jiang, J. Han, J. Liu, G. Zhang, L. Wang, F. Liu, X. Zhang, Y. Zhao, and D. Pang, "Risk of genome-wide association study newly identified genetic variants for breast cancer in Chinese women of Heilongjiang Province.," *Breast Cancer Research and Treatment*, vol. 128, no. 1, pp. 251–257, 2011.

[180] C. C. Teerlink *et al.*, "Genome-wide association of familial prostate cancer cases identifies evidence for a rare segregating haplotype at 8q24.21," *Human Genetics*, vol. 135, no. 8, pp. 923–938, 2016.

[181] S. Lindström *et al.*, "Common variants in ZNF365 are associated with both mammographic density and breast cancer risk," *Nature Genetics*, vol. 43, no. 3, pp. 185–187, 2011.

[182] R. Hamam *et al.*, "microRNA expression profiling on individual breast cancer patients identifies novel panel of circulating microRNA for early detection," *Scientific Reports*, vol. 6, p. 25997, 2016.

[183] A. R. Halvorse *et al.*, "Differential DNA methylation analysis of breast cancer reveals the impact of immune signaling in radiation therapy," *International Journal of Cancer*, vol. 135, no. 9, pp. 2085–2095, 2014.

[184] W. G. Jiang, G. Watkins, A. Douglas-Jones, L. Holmgren, and R. E. Mansel, "Angiomotin and angiomotin like proteins, their expression and correlation with angiogenesis and clinical outcome in human breast cancer," *BMC Cancer*, vol. 6, no. 1, p. 16, 2006.

[185] R. S. Redis *et al.*, "CCAT2, a novel long non-coding RNA in breast cancer: expression study and clinical correlations," *Oncotarget*, vol. 4, no. 10, pp. 1748–1762, 2013.

[186] M. M. W. Hijazi *et al.*, "NRG-3 in human breast cancers: activation of multiple erbB family proteins," *International Journal of Oncology*, vol. 13, no. 5, pp. 1061–1067, 1998.

[187] M. Mohseni *et al.*, "MACROD2 overexpression mediates estrogen independent growth and tamoxifen resistance in breast cancers," *Proceedings of the National Academy of Sciences*, vol. 111, no. 49, pp. 17606–17611, 2014.

[188] K. Michailidou *et al.*, "Large-scale genotyping identifies 41 new loci associated with breast cancer risk," *Nature Genetics*, vol. 45, no. 4, pp. 353–361, 2013.

[189] J. L. Caswell, R. Camarda, A. Y. Zhou, S. Huntsman, D. Hu, S. E. Brenner, N. Zaitlen, A. Goga, and E. Ziv, "Multiple breast cancer risk variants are associated with differential transcript isoform expression in tumors," *Human Molecular Genetics*, vol. 24, no. 25, pp. 7421–7431, 2015.

[190] Q. Zhu *et al.*, "Trans-ethnic follow-up of breast cancer GWAS hits using the preferential linkage disequilibrium approach," *Oncotarget*, vol. 7, no. 50, pp. 83160–83176, 2016.

[191] A. M. Dunning *et al.*, "Breast cancer risk variants at 6q25 display different phenotype associations and regulate ESR1, RMND1 and CCDC170.," *Nature Genetics*, vol. 48, no. 4, pp. 374–386, 2016.

[192] I. Barnett, R. Mukherjee, and X. Lin, "The generalized higher criticism for testing SNP-set effects in genetic association studies," *Journal of the American Statistical Association*, vol. 112, no. 517, pp. 64–76, 2017.