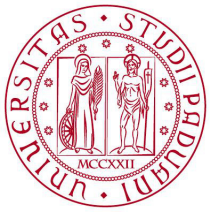


UNIVERSITÀ
DEGLI STUDI
DI PADOVA

Sede Amministrativa: Università degli Studi di Padova

Dipartimento di Scienze Statistiche
SCUOLA DI DOTTORATO DI RICERCA IN SCIENZE STATISTICHE
CICLO XXVI

Statistical Approaches in Genome-Wide Association Studies

Direttore della Scuola: Ch.ma Prof.ssa Monica Chiogna

Supervisore: Ch.ma Prof.ssa Monica Chiogna

Dottoranda: Akram Yazdani

31 January 2014

To My Parents

Acknowledgment

I owe my deepest gratitude to the head of the department Professor Alessandra Salvan who provided me the opportunity to join a research project in genome-wide association studies. This experience helped me to realize my enthusiasm for this field of research. I also express my warmest gratitude to my supervisor Professor Monica Chiogna for her support, vision and for considering my interest in Bayesian approach.

I am grateful to Professor Guido Massarotto. His guidance into the frequentist approach helped me to find my path in this area. Moreover, I want to give my gratitude to Professor Livio Finos who gave me valuable suggestions.

I would like to show my deepest gratitude to Professor David Dunson, Duke University, who gave me valuable comments and provided insight and direction-right up to the end. My gratitude is also extended to Doctor Mohammad Shariati, Ferdowsi University of Mashhad, who helped me understand better real genetic data set.

I am indebted to my colleagues and friends, Shireen Assaf, Md Abud Darda, Ivan Luciano Danesi, Lorenzo Maragoni, Roberta Pappadà, Luca Sartore, Erlis Ruli and also my other friends for making life enjoyable in Padova. I especially thank Luca Sartore because of his help in computer programming.

I gratefully acknowledge the funding source from Catholic University in Milan. I am also grateful to all the staff of Ph.D. school in Padova.

Last but not least, I would like to express my deepest gratitude to my family for all their love and encouragement. For my parents who raised me with a love of science and supported me in all my pursuits. For my siblings, Parisa, Hossein, Mandana, Mahdieh who inspired me with their moral support to continue my work with determination.

Akram Yazdani
University of Padova
January 2014

Abstract

Genome-wide association studies, GWAS, typically contain hundreds of thousands single nucleotide polymorphisms, SNPs, genotyped for few numbers of samples. The aim of these studies is to identify regions harboring SNPs or to predict the outcomes of interest. Since the number of predictors in the GWAS far exceeds the number of samples, it is impossible to analyze the data with classical statistical methods. In the current GWAS, the widely applied methods are based on single marker analysis that does assess association of each SNP with the complex traits independently. Because of the low power of this analysis for detecting true association, simultaneous analysis has recently received more attention. The new statistical methods for simultaneous analysis in high dimensional settings have a limitation of disparity between the number of predictors and the number of samples. Therefore, reducing the dimensionality of the set of SNPs is required.

This thesis reviews single marker analysis and simultaneous analysis with a focus on Bayesian methods. It addresses the weaknesses of these approaches with reference to recent literature and illustrating simulation studies. To bypass these problems, we first attempt to reduce dimension of the set of SNPs with random projection technique. Since this method does not improve the predictive performance of the model, we present a new two-stage approach that is a hybrid method of single and simultaneous analyses. This full Bayesian approach selects the most promising SNPs in the first stage by evaluating the impact of each marker independently. In the second stage, we develop a hierarchical Bayesian model to analyze the impact of selected markers simultaneously. The model that accounts for related samples places the local-global shrinkage prior on marker effects in order to shrink small effects to zero while keeping large effects relatively large. The prior specification on marker effects, which is hierarchical representation of generalized double Pareto, improves the predictive performance. Finally, we represent the result of real SNP-data analysis through single-maker study and the new two-stage approach.

Sommario

Lo Studio di Associazione *Genome-Wide*, GWAS, tipicamente comprende centinaia di migliaia di polimorfismi a singolo nucleotide, SNPs, genotipizzati per pochi campioni. L'obiettivo di tale studio consiste nell'individuare le regioni cruciali SNPs e prevedere gli esiti di una variabile risposta. Dal momento che il numero di predittori è di gran lunga superiore al numero di campioni, non è possibile condurre l'analisi dei dati con metodi statistici classici. GWAS attuali, i metodi negli maggiormente utilizzati si basano sull'analisi a marcatore unico, che valuta indipendentemente l'associazione di ogni SNP con i tratti complessi. A causa della bassa potenza dell'analisi a marcatore unico nel rilevamento delle associazioni reali, l'analisi simultanea ha recentemente ottenuto più attenzione. I recenti metodi per l'analisi simultanea nel multidimensionale hanno una limitazione sulla disparità tra il numero di predittori e il numero di campioni. Pertanto, è necessario ridurre la dimensionalità dell'insieme di SNPs.

Questa tesi fornisce una panoramica dell'analisi a marcatore singolo e dell'analisi simultanea, focalizzandosi su metodi Bayesiani. Vengono discussi i limiti di tali approcci in relazione ai GWAS, con riferimento alla letteratura recente e utilizzando studi di simulazione. Per superare tali problemi, si è cercato di ridurre la dimensione dell'insieme di SNPs con una tecnica a proiezione casuale. Poiché questo approccio non comporta miglioramenti nella accuratezza predittiva del modello, viene quindi proposto un approccio in due fasi, che risulta essere un metodo ibrido di analisi singola e simultanea. Tale approccio, completamente Bayesiano, seleziona gli SNPs più promettenti nella prima fase valutando l'impatto di ogni marcatore indipendentemente. Nella seconda fase, viene sviluppato un modello gerarchico Bayesiano per analizzare contemporaneamente l'impatto degli indicatori selezionati. Il modello che considera i campioni correlati pone una *priori* locale-globale ristretta sugli effetti dei marcatori. Tale prior riduce a zero gli effetti piccoli, mentre mantiene gli effetti più grandi relativamente grandi. Le priori specificate sugli effetti dei marcatori sono rappresentazioni gerarchiche della distribuzione Pareto doppia; queste

a priori migliorano le prestazioni predittive del modello. Infine, nella tesi vengono riportati i risultati dell'analisi su dati reali di SNP basate sullo studio a marcatore singolo e sul nuovo approccio a due stadi.

Contents

List of Figures

List of Tables

List of Abbreviations

BF	Bayes Factor
BL	Bayesian Lasso
BR	Bayesian Ridge
fdr	Local False Discovery rate
FDR	False Discovery Rate
FWER	Family Wise Error Rate
GDP	Generalized Double Pareto
GWAS	Genome-Wide Association Studies
HWE	Hardy-Weinberg Equilibrium
LD	Linkage Disequilibrium
MSPE	Mean Square Predictive Error
PCA	Principal Component Analysis
RP	Random Projection
SNP	Single Nucleotide Polymorphism

Chapter 1

Introduction

The ability of cost-efficient genotyping technologies brings the possibility of studying the relationship between complex traits or diseases with single nucleotide polymorphisms, SNPs, over entire genome. Genome-wide association studies, GWAS, usually include hundreds of thousands of SNPs assayed for few numbers of experimental units. The aims of studies are prediction or identifying regions harboring SNPs that affect the outcomes.

In the current GWAS, the widely applied methods are based on single marker analysis that does assess association of each SNP with the complex traits independently. However susceptibility loci have successfully identified from single based studies, the key problem of what threshold to use so as to select true association remains unresolved. An alternative approach is to analyze all SNPs simultaneously. This approach has recently received more attention by presence of new statistical methods appropriate for large scale problems. The main challenge with the use of these methods is the large disparity between the number of predictors, SNPs, and the number of observations in the model that reduces the accuracy of prediction and selection. Therefore, applying multi-stage analysis in the GWAS is required.

1.1 Overview

In the context of single marker analysis, p-value is the typical measure of statistical evidence of association between genetic variants and a complex trait of interest. The computed p-values for the null hypotheses of no associations lead to multiple hypotheses testing so as to identify the associated SNPs through multiple comparisons. Benjamini & Hochberg (1995) adopted the traditional multiple comparisons methods for large scale problems by introducing the concept of false discovery rate, FDR. The empirical Bayes version of FDR named local false discovery rate introduced by Efron *et al.* (2001). Then, Storey (2002) discussed the relation between these concepts. The posterior probability of association, PPA, is a full Bayesian approach for single marker analysis that can be thought of as the Bayesian analogue of the p-value (reviewed by Stephens & Balding, 2009).

For large-scale problems, $p \gg n$, in linear regression, there is a mass of literature in both frequentist and Bayesian framework. Frequentist imposes constraints on the size of coefficients, which can be seen as an extra term, known as penalization. The most popular one is L_1 norm penalty called the lasso (Tibshirani, 1996). The lasso has a parsimony property and also computational advantages via LARS algorithm (Efron *et al.*, 2004). Although the lasso is feasible from the point of view of computational complexity and selects the variables simultaneously, the rate of shrinkage is not desirable; it shrinks all coefficients with the same rate. A more desirable penalization is the one that strongly shrinks the small effects to zero and avoids shrinkage on the large effects. This can be achieved by imposing a concave penalization term into the regression model. Smoothly clipped absolute deviation penalty, SCAD, (Fan & Li, 2001) and minimax concave penalty (Zhang, 2010) can be named as concave penalization methods.

To deal with the complexity due to $p \gg n$, one of the Bayesian approaches is to consider a mixture prior. A point mass mixture prior is specific form of this class that is widely applied in variable selection or model selection contexts. One of the early methods based on mixture prior is the stochastic Bayes variable selection proposed by George & McCulloch (1993). This kind

of prior correctly represents sparsity assumption by placing positive mass at zero. The optimal properties can be also achieved by carefully choosing point mass mixture prior (Castillo & van der Vaart, 2012). However this approach gets popularity in different applications as well as in genetic, it is not efficient for problems like GWAS since its computational complexity is exponential in the number of predictors.

Another Bayesian approach is based on the global-local shrinkage prior that models the regression coefficients with absolutely continuous shrinkage prior at zero. Such a prior is computationally attractive and also capable for nearly sparse problems. Hence, a large number of literatures is devoted to present new types of shrinkage priors and discusses their properties. Here, we just refer to some of those, Armagan *et al.* (2013), Carvalho *et al.* (2009), Park & Casella (2008) and Griffin & Brown (2007). Although global-local shrinkage priors provide some advantages like computational efficiency, they create their own challenges because the posterior probability mass on a regression coefficient equal to zero is never positive.

However the aforementioned approaches for simultaneous analysis have been introduced for $p \gg n$, the large disparity between p and n in the GWAS causes a poor predictive performance. Thus before any analysis, there is a need to reduce dimension of the parameter space. One of the dimensional reduction techniques is random projection. The idea is based on projecting data in low dimensional space randomly while preserving the distances between points. This ensures that we can learn from projected data about the response with little loss of information. Various literatures have discussed the accuracy of random projection by introducing a boundary on the size of new space (see, e.g., Dasgupta & Gupta, 2003; Achlioptas, 2003; Li *et al.*, 2006). Typically, random projection has been studied from two points of view. One idea is to use random projections to compress the samples and the other one is random projections on the parameter space. The latter can be related to the problem in the GWAS. In the context of linear regression, Maillard & Munos (2009) shows a bias-variance trade-off with assumption of i.i.d. samples. Fard *et al.* (2012) provides a bias-variance analysis of ordinary

least-squares regression in compressed spaces with sparsity assumption. It shows that the sparsity assumption allows working with non i.i.d. samples. Guhaniyogi & Dunson (2013) introduces Bayesian compressed regression that shows good performance for dense problems.

Two-stage approaches can be an alternative to improve accuracy of aforementioned statistical approaches in the GWAS. The task in the first-stage is to screen all markers in order to select the most promising markers. This provides a small set of predictors appropriate for simultaneous analysis in the second-stage. Fan & Lv (2008) and Paul *et al.* (2008) propose two-stage procedure for variable selection. Li *et al.* (2011) integrates Paul et al.'s first-stage procedure into Bayesian Lasso for identifying important SNPs.

1.2 Main contributions of the thesis

The focus of the present thesis is to overcome complexity in high dimensional settings similar to the GWAS in order to provide an accurate prediction. Motivation of this study is a Genome-wide problem in animal breeding which genotyped about 707,962 SNPs for 607 Holstein Bulls. The purpose is to improve milk productivity through investigating protein yield and longevity phenotype.

Chapter 2 briefly explores single marker analysis. We first consider multiple hypotheses testing for large scale problems. To adjust multiple comparisons for these kinds of problems, the false discovery rate and the local false discovery rate are reviewed. Then, we explain how to select associated SNPs via Bayes factor with reference to recent studies. We also look at the relation between the Bayes factor and standard frequentist hypothesis testing.

Chapter 3 first looks in on penalization approaches and then explores Bayesian methods appropriate for large p and small n. However such methods are widely applied, unfortunately, they face multi-layered challenges in genome-wide problems. These challenges can be addressed to efficiency and accuracy of the result. To find a better picture about the predictive performance of these methods for the large disparity between p and n, we have

illustrated a simulation study. The result of the simulation study confirms that the dimension of the parameter space in the problems like the GWAS must be reduced before main analysis.

In Chapter 4, we focus on random projection techniques to suppress predictors into a low dimensional space. Since our analysis is based on a linear mixed model, we modified Bayesian compress regression by Guhaniyogi & Dunson (2013) introduced for linear regression. To evaluate the performance of this approach for SNP-data, a simulation study has been conducted and the result has been compared with predictive performance of un-projected data.

In Chapter 5, we present a new two-stage approach for problems with related samples such as family studies. This approach is a hybrid method of single marker analysis and simultaneous analysis. In the first-stage, we list markers by the posterior odds of presence of each SNP in the model at a time. For selecting the most promising SNPs in this stage, we consider two different thresholds. One is defined as a typical threshold in single marker analysis that provides possibility to consider epistatic effects in the model for simultaneous analysis. The other one is equivalent to safe upper limit of the number of predictor in the second-stage model. With this choice of threshold that reduces the risk of missing important SNPs, the second-stage model includes the marginal effects. In the second-stage, we develop two models corresponding two the different threshold. In the both linear mixed models, we implement generalized double Pareto as shrinkage prior (Armagan *et al.*, 2013) on marker effects. With these prior specifications, we estimate parameters of the models by sampling from their conditional posterior distributions through the MCMC algorithm.

The last chapter is devoted to application to the real genome-wide association problem. After introducing the problem and some preliminary analyses, we have attempted to identify true genetic association through the multiple hypotheses testing reviewed in Chapter 2. Then, predictive performance of the proposed method in Chapter 5 has been evaluated with 10-fold cross validation and also comparison with two other prior specifications. We then

selected SNPs based on the heritability through fitting the model over whole samples. As the result, 32 SNPs have been selected as the most promising markers. Then, we have applied the model with epistatic effects for selected SNPs and calculated the total genetic variance contributed by the marginal and epistatic effects.

Chapter 2

Single Marker Analysis

In genome-wide association studies, the first attempts to incorporate marker information into prediction and identify region harboring SNPs were based on single marker analysis. This analysis for quantitative traits have been illustrated through linear regression. However fitting a linear regression on a single marker is simple; the main challenge in this context is how to define the threshold for detecting a subset of SNPs truly associated with the traits.

In this chapter, we first consider multiple hypotheses testing for large scale problems. To adjust for multiple comparisons of the large number of hypotheses, we focus on false discovery rate and its empirical Bayesian version named local false discovery rate. We then look at full Bayesian approach particularly in its application in the GWAS. Finally, the relation between the full Bayesian approach and standard frequentist approach in the GWAS is presented.

2.1 Multiple Hypothesis Testing

A widely used approach to identify significant association is to analyze one SNP at a time that is based on univariate linear regression for quantitative traits. Fitting a linear regression for each SNP at a time leads to test a large number of hypotheses. If P denotes the number of SNPs that contribute to

the analysis, we have a set of hypotheses as

$$\begin{cases} H_{0i} : \beta_i \text{ is not significant}, & i = 1, ..., P \\ H_{1i} : \beta_i \text{ is significant}, & i = 1, ..., P. \end{cases}$$

Since many hypotheses are tested simultaneously, a multiple comparisons procedure needs to be applied in order to avoid spurious detection. The family wise error rate, FWER, is the classical approach that controls the overall Type I error at level α. Following the notation in Table 2.1, FWER is defined as

$$\text{FWER} = p(a \geq 1),$$

which is the probability of making one or more false positive discovery among all the hypotheses.

		Decition		
		Null	Non-Null	
True	Null	$P_0 - a$	a	P_0
	Non-Null	$P_1 - b$	b	P_1
		$P - R$	R	P

Table 2.1: multiple hypotheses testing

While in the GWAS P is too large, setting a threshold based on FWER is too strict and prevents detecting SNPs associated with the traits. Hence, many studies have been focused on adjusting multiple comparisons to large scale problems like genetic problems (see, e.g.,Dudbridge & Gusnato, 2008; Goemana & Aldo Solarib, 2014; and references therein).

2.1.1 False Discovery Rate

In large scale problems with tens of thousands of hypotheses, controlling Type I error might not provide a good threshold since it is corresponding with

low power of detecting significant association. To overcome this problem, Benjamini & Hochberg (1995) introduced false discovery rate, FDR, that increases the power by tolerating some Type I errors. The FDR controls the proportion of false discovery, a, to the total discovery, R, as

$$\text{FDR} = \text{E}\left(\frac{a}{R \vee 1}\right) \tag{2.1}$$

where $R \vee 1 \equiv \max(R, 1)$. However the most obvious definition of a false discovery rate is $E\left(a/R\right)$, the FDR in (2.1) is a remedy to prevent undefined situation in the case that $R = 0$. An alternative can be positive false discovery rate suggested by Storey (2003) as

$$pFDR = E\left(\frac{a}{R}\middle| R > 0\right).$$

Since P is too large in the GWAS, the probability of $R > 1$ is almost 1 and both above quantities are approximately equal. Hence, in this context, the FDR can be estimated simply through $E(a)/E(R)$ for a specific threshold.

The FDR offers less stringent control over Type I errors than the FWER; therefore, control of the FDR is close to a weak control of the FWER (Benjamini & Hochberg, 1995). For instance in the case that all null hypotheses are true, i.e. $a = R$, then

$$\begin{cases} \text{if } a \geq 1 \Rightarrow & \dfrac{a}{R} = 1, \\ \\ \text{if } a = 0 \Rightarrow & \dfrac{a}{R} = 0, \end{cases}$$

implies

$$\text{FDR} = \text{E}\left(\frac{a}{R}\right) = p(a \geq 1) = FWER.$$

Therefore, we can say the FWER is upper bound of the FDR and controlling the FWER is equivalent to controlling the FDR.

In practice for selecting significant SNPs based on the FDR, the p-values for all test statistics are required. After calculating all p-values, we need to rank them such that $p_{(1)} \leq p_{(2)} \leq ... \leq p_{(P)}$. We then reject the hypotheses with p-values under the $p_{(k)}$ where

$$k = \max\left[i \in \{1, ..., P\} : p_{(i)} \leq \frac{iq}{P}\right] \tag{2.2}$$

and q is a fixed value in $(0, 1)$. If no p-value satisfies inequality (2.2), then no hypothesis test is called significant. The above procedure that is so called BH-algorithm controls the FDR at level

$$FDR = q \times \left(\frac{P_0}{P}\right) \leq q$$

if the statistical tests are independent. The main problem with the use of FDR in the GWAS is the independency assumption that cannot be fulfilled due to the linkage disequilibrium among SNPs. Although Benjamini & Yekutieli (2001) weaken the independence condition to positive regression dependence, this condition does not hold in the GWAS as well.

2.1.2 Local False Discovery Rate

Local false discovery rate, fdr, (Efron *et al.*, 2001) is an empirical Bayesian version of the false discovery rate. The main assumption underling the theory behind the fdr is to assume that each statistic probabilistically follows a random mixture of a null distribution and non-null distribution; it is the main assumption in some literature (see, e.g., Lee *et al.*, 2000; Newton *et al.*, 2001; Storey, 2003; Storey & Tibshirani, 2003). To define the fdr, each test statistics t_i requires being converted to z-value as

$$z_i(x) = \Phi^{-1}(G_0(t_i))$$

where Φ^{-1} is the inverse function of the standard normal cumulative density function, cdf, and G_0 is a putative null cdf of the t_i. The use of z-values makes analysis more convenient due to the properties of normal theory. With mixture distribution assumption, marginal density of each z-value is

$$f(z) = \pi_0 f_0(z) + \pi_1 f_1(z),$$

where

$$\begin{cases} f_0(z) = f(z \mid \text{null}) & \text{with} \quad p(\text{null}) = \pi_0 \\ f_1(z) = f(z \mid \text{non-null}) & \text{with} \quad p(\text{non-null}) = \pi_1. \end{cases}$$

The fdr is then the posterior probability of null case given the z-value:

$$fdr(z) \equiv p(\text{null}|z) = \pi_0 f_0(z)/f(z). \tag{2.3}$$

In order to illustrate the relationship between the fdr and the FDR, we need to focus on tail areas rather than densities while the FDR relies on. Thus, let $F_0(z)$ and $F_1(z)$ denote the cdf's corresponding to $f_0(z)$ and $f_1(z)$ and define $F(z) = p_0 F_0(z) + p_1 F_1(z)$, then

$$\text{Fdr}(z) \equiv p(\text{null} \mid Z \leq z) = \pi_0 F_0(z)/F(z). \tag{2.4}$$

The (2.4) implies that Benjamini and Hochberg's FDR control rule depends on an estimated version of (2.4) where F is replaced by the empirical cdf. The fdr that is a Bayesian approach offers some insight to define the cutoff threshold through posterior odds ratio

$$p(\text{non-null} \mid z)/p(\text{null} \mid z) = (1 - \text{fdr})/(\text{fdr})$$

while there is no consensus on a standard choice of q for the FDR in (2.2).

Bayesian false discovery rates, both the fdr and the Fdr, depend on the marginal distribution of the z-values, $f(z)$ or $F(z)$. On one side, assumption of independent z-values is not required despite assumption of independency of p-values for the FDR. On the other side, the inference is based on analysis of one SNP at a time, this may be quite different from the posterior probability of H_{i0} given entire P vector of z-values.

As it is clear from (2.3), the fdr does not directly depend on $f_1(z)$. i.e., the density of non-null cases is not required for estimating the fdr. To estimate the numerator in (2.3), it might be assumed that $f_0(z)$ follows the theoretical null density which is $N(0, 1)$; however in large scale problems, the empirical density is usually wider (thinner) than theoretical null density. Efron (2004, 2007) discussed the reasons why f_0 might differ from the theoretical null. In these kinds of cases, one might consider the theoretical null density as $N(\mu_0, \sigma_0^2)$ instead of $N(0, 1)$. With this assumption, the parameters of theoretical null density need to be estimated.

2.1.3 Estimating $\pi_0 f_0(z)$

To estimate $\pi_0 f_0(z)$, let assume A_0 is a subset of sample space near zero such that

$$z \in A_0 \Rightarrow f_1(z) = 0. \tag{2.5}$$

This implies

$$z \in A_0 \Rightarrow f(z) = \pi_0 f_0(z).$$

Hence, when $z \in A_0$, we have

$$\log(f(z)) = \left[\log(\pi_0) - \frac{1}{2}\left(\frac{\mu_0^2}{\sigma_0^2} + \log(2\pi\sigma_0^2)\right)\right] + \frac{\mu_0}{\sigma_0^2}z - \frac{1}{2\sigma_0^2}z^2. \tag{2.6}$$

The parameters (π_0, μ_0, σ_0) can be estimated through *maximum likelihood* or *central matching* methods (Efron, 2004).

- Central Matching Approach
  In the central matching, we assume $\log(f(z))$ is a quadratic function around zero as
  $$\log(f(z)) \doteq \gamma_0 + \gamma_1 z + \gamma_2 z^2.$$
  To estimate γ's, we partition the range Z into K bins, Z_k, with width of Δ such that
  $$Z = \cup_{k=1}^{K} Z_k.$$
  Then, we define a count variable y_k as
  $$y_k = \#\left[z_i \in k\text{th bin}\right], \quad k = 1, 2, ..., K,$$
  that is order statistic of z when $\Delta \to 0$. By estimating γ_is from the histogram counts of y_k around $z = 0$ and matching with coefficients in (2.6), the estimate of parameters of null density can be obtained.

- Maximum Likelihood Method
  To estimate parameters (μ_0, σ_0, π_0) by maximum likelihood method, the joint density of z-values in A_0 should be obtain. To this end, let define
  $$I_0 = \{i : z_i \in A_0\} \text{ and } P_0 = \#I_0.$$

Then, for z-values belong to A_0

$$f(z) = \left[\theta^{P_0}(1-\theta)^{P-P_0}\right] \left[\prod_{I_0} \frac{\varphi_{\mu_0,\sigma_0}(z_i)}{H_0(\mu_0,\sigma_0)}\right]$$

where φ is density function of $N(\mu_0, \sigma_0)$, $H_0(\mu_0, \sigma_0) \equiv \int_{A_0} \varphi_{\mu_0,\sigma_0}(z_i)dz$ and $\theta = \pi_0 H_0(\mu_0, \sigma_0) = p(z_i \in A_0)$. This method yields smaller variation but more bias for the estimators.

Both aforementioned approaches rely on the assumption (2.5) that may not hold in practice for all z-values in A_0 . This introduces some bias to the estimator, but it is can be ignored if π_0 is close to one.

Although we have possibility to estimate theoretical null density in large-scale problems, the difference between empirical density and the theoretical null density might be due to a kind of structure in the data. In genetic problems, it is very common to have population stratification or related samples. Ignoring these kinds of structures may be the cause of this difference. For instance in our experience presented in Chapter 6, by adding random effect to the model and accounting for related samples the theoretical null density, $N(0,\ 1)$, turns to be true for our SNP-data.

2.2 Full Bayesian Approach

Bayesian methods provide an alternative approach to p-value by computing posterior probability of association, PPA, that is defined as

$$PPA = PO/(1 + PO). \tag{2.7}$$

The PO is posterior odd of model with single SNP, M_1, against the model without any marker effects, M_0:

$$\begin{aligned} \text{PO} &= \left(\frac{ML_1}{ML_0}\right) \times \left(\frac{\pi}{1-\pi}\right) \\ &= \quad BF \quad \times \quad \text{PriorOdd}. \end{aligned}$$

The probability π quantifies our belief in association of SNP with the complex trait. The Byes factor, BF, is the ratio of the marginal likelihood of the model M_1 to the model M_0; i.e., it measures the consistency of the set of data with a non null hypothesis in comparison with the null. Therefore, the PO and consequently the PPA incorporate the prior knowledge in making decision via evidence from data. The prior knowledge that represent in the model through π can be varied across SNPs. However, if π is assumed to be the same for all SNPs, it can be interpreted as a prior estimate of the overall proportion of SNPs that are truly associated with the phenotype. In this case, the comparison among SNPs can be done via the BF.

2.2.1 The Choice of Threshold for the BF

Since the PPA can be easily obtained from the BF given π, the BF is often used as the primary summary of the evidence for association at a SNP. In many applications, a typical threshold for BF is 10, which is corresponding to strong evidence against null hypothesis (Jeffreys, 1961). In contrast, this number cannot be an appropriate threshold in GWAS since it does not provide high posterior probability of association (Wakefield, 2007; Stephens & Balding, 2009).

While in single marker analysis the aim is to select the most promising SNP, it is assumed that a minority of SNPs is expected to be truly associated with the phenotype (e.g., Diabetes Genetics Initiative of Broad Institute of Harvard and MIT, 2007; Wellcome Trust Case Control Consortium, 2007). Due to this assumption, the suggested range of prior probability π is in $(10^{-4}, 10^{-6})$ (Stephens & Balding, 2009; Ball, 2011). The use of this small prior probability of true association provides a very small prior odd. Therefore, in order to have the PPA close to one, which is correspond to high posterior odd, the BF is required to be big enough to overcome low prior odd. For instance, if $\pi = 10^{-4}$, the BF greater than 10^4 is required to provide a PPA close to one.

2.3 Relation between the BF and p-value

In the Bayesian approach, we define a threshold on the Bayes factor that seems strange in compare to many other applications. The requirement for a large BF is analogous to setting a stringent threshold for the GWAS in a frequentist approach. For clarity of this claim, let define the posterior odd based on a quasi-Bayesian argument for a class of tests with $T > t$ as significant statistics :

$$\frac{p(H_1 \mid T > t)}{p(H_0 \mid T > t)} = \frac{p(T > t \mid H_1)}{p(T > t \mid H_0)} \times \frac{p(H_1)}{p(H_0)}$$

$$= \frac{1-\beta}{\alpha} \times \frac{p(H_1)}{p(H_0)}$$

where α and β are the Type I and Type II error rates (Wellcome Trust Case Control Consortium, 2007); hence, with this representation, the Bayes factor is a function of power and Type I error. Assuming a problem with 10^6 independent markers with 10 SNPs associated with the trait and average power 50% to detect an associated SNP. In order to achieve the posterior odds of $10:1$ in favor of association, a p-value of 5×10^{-7} is required (Dudbridge & Gusnato, 2008); therefore,

$$\frac{1-\beta}{\alpha} = 10^6.$$

In addition to the above argument that make a bridge between Bayes factor and p-value, experience in the GWAS confirms that p-values and Bayes factors would give the same ranking of SNPs in order of strength of evidence for well defined test statistics; although they are different in terms of interpretation and statistic value. For instance in Bayesian analysis, the large BF is required due to prior belief not the large number of tests that are actually or potentially performed.

2.4 Discussion

The single SNP-based studies in genome-wide problems have been instrumental in detecting significant genes for various complex diseases or traits. These approaches may measure the evidence of association through p-value, z-value or Bayes factor. Typical threshold for selecting significant association must be adjusted to the GWAS. The false discovery rate and its empirical Bayes version so called local false discovery rate adjust the cutoff threshold on p-vales and z-values by controlling the rate of false positive discovery. In full Bayesian approach, the threshold for Bayes factor should be redefined due to the prior belief that a few numbers of SNPs are associated with the trait. However susceptibility loci have successfully identified by these adjustments, single marker analysis may not be powerful for identifying weaker associations and also cannot consider the epistatic effects.

Chapter 3

Simultaneous Analysis

Better understanding of biological system requires considering all markers simultaneously in the model. This makes the model capable of explaining phenotipic variance and consequently predicting quantitative traits or disease susceptibility of future individuals. The main challenge for this kind of studies is the large number of markers in the model. Typically, in genome-wide association studies the number of markers, p, vastly exceeds the number of observations, n, that breaks down the main assumption in classical methods. To deal with $p \gg n$ problem, penalization or thresholding methods have been introduced in the frequentist context. On the other hand, Bayesian approaches attempt to overcome this difficulty by specifying new form of priors. These priors can be divided into two main categories, shrinkage priors and mixture priors. Shrinkage priors are continuous priors concentrated at zero in order to shrink marker effects toward the origin. The rate of shrinkage that is controlled by hyperparameters of the priors should be adjusted automatically with the effect sizes, i.e, the magnitude of small effects toward zero should be stranger than the one for large effects. Another prior specification in high dimensional settings is based on discrete mixture of distributions. The main assumption of mixture priors is that set of markers is a collection of some set of markers with different patterns for size of effect. A widely applied mixture prior is mixture of set of zero and nonzero effect sizes.

Here, we consider a normal linear regression model

$$\mathbf{y} = \mathbf{x}_1\beta_1 + ... + \mathbf{x}_p\beta_p + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim N(0, \sigma^2 I_n) \tag{3.1}$$

where $\mathbf{x}_i$s are n-vectors of genotyped markers, SNP, β_is are marker effects. We also assume that the $\mathbf{x}_i$s are centered.

While most of the point penalization estimates of β_is correspond to the mode of a posterior distribution obtained under shrinkage priors, we first briefly explore some penalization methods. Then we look at the shrinking concept in the Bayesian framework by making a connection between Bayesian and frequentist approach. Section 3 and 4 are devoted to shrinkage and mixture priors. In section 5, we evaluate performance of these methods through simulation studies for different number of predictors in the model.

3.1 Penalized Method

While classical methods like maximum likelihood estimation break down in $p \gg n$ problems, some constraints are required on the size of effects. The methods which impose some restriction to the model are known as penalized methods with point estimate of $\boldsymbol{\beta}$ as

$$\hat{\boldsymbol{\beta}} = \arg\min_{\beta} \big(\parallel \mathbf{y} - X\boldsymbol{\beta} \parallel_2^2 + p_\lambda(\boldsymbol{\beta}) \big) \tag{3.2}$$

where $p_\lambda(\boldsymbol{\beta})$ is the penalized function and $\lambda > 0$ is the penalty (tuning) parameter. This estimate can be thought as shrunken least square estimator. The rate of shrinkage is related to defined penalized function. Hence, different penalized functions have been presented in literature in order to improve and adopt them for different problems. Here we just explore some of the most popular methods.

3.1.1 The Lasso and Adaptive Lasso

The most popular and widely used penalization method is the lasso with penalized function $\lambda \parallel \boldsymbol{\beta} \parallel_1 = \lambda \sum_i \mid \beta_i \mid$ (Tibshirani, 1996). Applying L_1

norm penalty instead of L_2 norm in ridge regression provides the parsimony property for the lasso. This kind of penalization simultaneously selects a subset of predictors as effective variables and shrinks the rest exactly to zero. Hence, L_1 penalty makes the lasso a continuous subset selection.

The (3.2) is an optimization problem of a convex function with lasso penalization. Hence, based on Karush-Kuhn-Tucker (KKT) condition for the global minimization, a necessary and sufficient condition for $\hat{\boldsymbol{\beta}}$ to be a solution of the lasso is

$$\begin{cases} G_i(\hat{\boldsymbol{\beta}}) = -\text{sign}(\hat{\beta}_i)\lambda & \text{if} \quad \hat{\beta}_i \neq 0 \\ \\ \mid G_i(\hat{\boldsymbol{\beta}}) \mid \leqslant \lambda & \text{if} \quad \hat{\beta}_i = 0 \end{cases}$$

where $G(\hat{\boldsymbol{\beta}}) = -2X^T(Y - X\boldsymbol{\beta})/n$. Moreover, if the solution is not unique and $G_i(\hat{\boldsymbol{\beta}}) < \lambda$ for some solution $\hat{\boldsymbol{\beta}}$, then $\hat{\beta}_i = 0$ for all solutions.

Although the lasso is feasible from the computational point of view and selects variables simultaneously, it does not have oracle properties unless it fulfills irrepresentable condition defined in the follow.

Neighborhood Stability and Irrepresentable Condition

The neighborhood stability condition is equivalent to the so called irrepresentable condition (Zou, 2006; Zhao & Yu, 2006). If we denote $\hat{\boldsymbol{\Sigma}} = n^{-1}\mathbf{X}^T\mathbf{X}$ and $S_0 = \{i; \beta_i^0 \neq 0\} = \{1, 2, ..., s_0\}$ which consists of the first s_0 variables, we can partition $\hat{\boldsymbol{\Sigma}}$ as

$$\hat{\boldsymbol{\Sigma}} = \begin{bmatrix} \hat{\boldsymbol{\Sigma}}_{1,1} & \hat{\boldsymbol{\Sigma}}_{1,2} \\ \\ \hat{\boldsymbol{\Sigma}}_{2,1} & \hat{\boldsymbol{\Sigma}}_{2,2} \end{bmatrix}$$

where $\hat{\boldsymbol{\Sigma}}_{1,1}$ is a $s_0 \times s_0$ matrix corresponding to the active variables,$\hat{\boldsymbol{\Sigma}}_{1,2} = \hat{\boldsymbol{\Sigma}}_{2,1}^T$ is a $s_0 \times (p - s_0)$ matrix and $\hat{\boldsymbol{\Sigma}}_{2,2}$ is a $(p - s_0) \times (p - s_0)$ matrix. Then the irrepresentable condition is

$$\|\hat{\boldsymbol{\Sigma}}_{2,1}\hat{\boldsymbol{\Sigma}}_{1,1}^{-1}\text{sign}(\beta_1^0, ..., \beta_p^0)\|_\infty \leq \theta \quad \text{for some} \quad 0 < \theta < 1,$$

where $\|x\|_\infty = \max_i |x^{(i)}|$ and $\text{sign}(\beta_1^0, ..., \beta_p^0) = (\text{sign}(\beta_1^0), ..., \text{sign}(\beta_p^0))^T$.

Having the upper bound $\theta < 1$ requires penalized (tuning) parameter $\lambda = \lambda_n$ to be chosen of a order larger than $\sqrt{\log(p)/n}$. Therefore, if the design matrix is too much ill posed and exhibits a strong degree of linear dependence within smaller sub-matrices of $\mathbf{X}$, the lasso performance will be poor and inconsistent. Actually for consistency of lasso, a strong irrepresentable condition on the covariance matrix $\mathbf{X}^T\mathbf{X}$ and some additional regularity conditions on $\{n, p, \beta\}$ must hold, which are not so practical. Moreover, the lasso estimator may even violate the sign consistency that causes a converse interpretation. Therefore, applying lasso for SNP-data may not provide an accurate result since we usually have highly correlated SNP.

In order to correct the overestimation behavior of the lasso, Zou (2006) introduces the adaptive lasso. He replaces L_1 penalty by a re-weighted version as

$$\sum_{i=1}^{p} |\beta| / |\hat{\beta}_{\text{init},i}|,$$

where $\hat{\beta}_{\text{init},i}$ is an initial estimator.
This method is a two-stage procedure. By cleverly choosing the weight, the adaptive lasso shows oracle properties. One choice of weights is based on a root-n consistent estimator $\hat{\beta}$ of β, for example the ML estimator when $p < n$. In high dimensional problem, in the first stage $\hat{\beta}_{\text{init}} = \hat{\beta}(\hat{\lambda}_{\text{init},CV})$ is estimated initially from the lasso, since the tuning parameter $\hat{\lambda}_{\text{init},CV}$ is estimated by cross validation. In the second stage, we use cross validation to select λ in adaptive lasso. Then we can expect that

- if $\hat{\beta}_{\text{adapt},i} = 0 \quad \Rightarrow \hat{\beta}_{\text{adapt},i} = 0$, and

- if $\mid \hat{\beta}_{\text{init},i} \mid$ is large, the adaptive lasso employs little shrinkage which provides less bias.

3.1.2 Minimax Concave Penalty, MCP

The minimax concave plus, MCP, is a penalization method that imposes concave penalized function to the model as

$$p_\lambda(\beta) = \lambda \int_0^\beta (1 - x/(\gamma\lambda))_+ dx$$

with a regularization parameter γ. The main idea behind the MCP is to shrink only the β_is which are small. This can be recognized by looking at the rate of the penalized function. The MCP shrinks the variables under threshold $\lambda\gamma$ with a shrinkage rate that is decreasing with the size of the β_is. Therefore, in MCP we have an unbiased estimator for coefficients above threshold and a shrunken estimator for the ones under threshold.

Since the MCP function is concave in order to have sparse convex penalized loss function, the convexity of loss function must overcome the convexity of MCP. This can be fulfilled by considering sparse Rize condition (SRC).

Sparse Riesz condition, SRC

Sparse Riesz condition on design matrix X for suitable $0 < c_* \leqslant c^* < \infty$ and rank d^* is given by

$$c_* \leq \min_{|S_0| \leq d^*} c_{\min}(\mathbf{\Sigma}_{1,1}) \leq \max_{|s_0| \leq d^*} c_{\max}(\mathbf{\Sigma}_{1,1}) \leq c^*,$$

where S_0, $\mathbf{\Sigma}_{1,1}$ have the same definition in definition of irrepresentable condition and $c_{\min\,/\,\max}(M)$ is the smallest/largest eigenvalue of M.

3.1.3 Elastic Net

Although, the above penalized methods improve the accuracy of prediction and minimize the residual sum of squares error, they are not appropriate methods when predictors are highly correlated. Typically, this kind of data have a group structure that predictors in each group are highly correlated. Therefore, this information should be taken into account for imposing constraints to the model.

The Elastic net is one of the penalization methods (Zou & Hastie, 2005) that is defined for these kind of data as

$$p_\lambda(\boldsymbol{\beta}) = \lambda_2 \parallel \boldsymbol{\beta} \parallel_2^2 + \lambda_1 \parallel \boldsymbol{\beta} \parallel_1 .$$

This optimization problem is a convex function. Elastic net shrinks in the direction of ridge regression by a lasso-type threshold. So it has characteristics of both, with the advantage of convexity that makes it useful for correlated data. It exhibits that correlated variables in the same group tend to have equal coefficients and the upper bound for difference of those coefficients is a function of the sample correlation as

$$\frac{1}{\|\mathbf{y}\|_1} \mid \hat{\beta}_i - \hat{\beta}_j \mid \leq \frac{1}{\lambda_2}\sqrt{2(1-\rho)},$$

where $\rho = \mathbf{x}_i^T \mathbf{x}_j$ the sample correlation.

Elastic net also overcomes a limitation of the lasso for the number of selected predictors which is at most as equal to the number of observations. This property comes from the idea of solving problem for augmented data. By adding artificial data set, we increase the rank of design matrix up to p, i.e., elastic net can potentially select all p predictors in all situation.

However elastic net has good properties, it does not perform satisfactorily unless it is very close to Ridge regression or the lasso. The weakness arises by the double shrinkage, first estimating the ridge coefficient and then the lasso type shrinkage. Shrinking twice does not reduce variance much and introduces more bias into the model, in comparison with ridge regression and the lasso. In order to undo the extra shrinkage, the estimate should be rescaled with $(1 + \lambda_2)$.

3.2 Posterior Expectation and Least Square Estimate

All penalized methods shrink the standard least square estimator, $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$, toward the origin. The shrinkage rate depends on the form of

penalized function. To make a connection between the shrinkage concept in Bayesian and frequentist framework, let's define the prior predictive of $\hat{\boldsymbol{\beta}}$ as

$$h(\hat{\boldsymbol{\beta}}) = \int L(\boldsymbol{\beta}; \hat{\boldsymbol{\beta}}) \pi(\boldsymbol{\beta}) d\boldsymbol{\beta}$$

where $L(\boldsymbol{\beta}; \hat{\boldsymbol{\beta}})$ denotes the likelihood function and $\pi(\boldsymbol{\beta})$ denotes a specified prior on $\boldsymbol{\beta}$. Following the density of $\hat{\beta} \sim N\left(\beta, \sigma^2 (X^T X)^{-1}\right)$, for $p < n$[1] we have

$$E(\boldsymbol{\beta}|\hat{\boldsymbol{\beta}}) = (I - S(\hat{\boldsymbol{\beta}}))\hat{\boldsymbol{\beta}}. \tag{3.3}$$

Here, $S(\hat{\boldsymbol{\beta}}) = \sigma^2 (X^T X)^{-1} R(\hat{\boldsymbol{\beta}})$ and $R(x)$ is a diagonal matrix with

$$R_{ii}(x) = -\frac{1}{x_i} \frac{\partial}{\partial x_i} \log h(x).$$

The (3.3) representation of the posterior expectation given by Griffin & Brown (2010) makes the comparison between the standard least square and posterior expectation of $\boldsymbol{\beta}$ easy. As it is clear, the posterior expectation is a shrunken version of $\hat{\boldsymbol{\beta}}$. The rate of shrinkage is controlled by the predictive prior and variance of $\hat{\boldsymbol{\beta}}$.

In the case of orthogonal designs, the posterior expectation can be simply expressed for each β_i as

$$E(\beta_i|\hat{\beta}_i) = \hat{\beta}_i (I - S^{(i)}(\hat{\beta}_i)),$$

where

$$S^{(i)}(\hat{\beta}_i) = \frac{\sigma^2}{\sum_j^n x_{ji}^2} R_{(ii)}(\hat{\beta}_i), \quad R_{(ii)}(\hat{\beta}_i) = -\frac{1}{\hat{\beta}_i} \frac{\partial}{\partial \hat{\beta}_i} \log h(\hat{\beta}_i).$$

It is clear that the shape of $\pi(\beta_i)$ is affected by the rate of shrinkage. For instance, if we place a normal prior on β_j, $h(\hat{\beta}_i)$ has normal tails such that

$$h(\hat{\beta}_i) \approx \exp(-\frac{1}{2} c \hat{\beta}_i^2) \quad \Rightarrow \quad R_{ii}(\beta_i) \to c.$$

This leads to undesirable shrinkage because $E(\beta_i|\hat{\beta}_i)$ does not limit to $\hat{\beta}_i$ as $\hat{\beta}_i \to \infty$; i.e. this choice of prior does not provide tail robustness property

[1] In the case that $p \geq n$, X is singular. Therefore singular value decomposition techniques should be utilized to extend the result.

which is discussed in section 3.3.5. To avoid this situation, a prior distribution with heavier tails than normal should be given to β_i (David, 1973). While the natural class of prior density for β_is in linear regression model is scale mixtures of normal, priors with heavier tails can be placed via the hierarchical form that is discussed in the next section.

3.3 Shrinkage Prior

Shrinkage priors are usually a continuous shrinkage with hierarchy representation. Hierarchical models conceptually and practically are at the center of attention in modern Bayesian statistics. On the theoretical side, hierarchical models allow a more objective approach to the inference by estimating hyperparameters from data rather than subjective approach (see, e. g,Efron & Moris, 1975). These models practically are more flexible tools for combining information and partial pooling of inference (see, e. g., Carlin & Louis, 2001; Gelman, 2003). The continuous shrinkage property is also an important characteristic since it avoids instability in model prediction (Fan & Li, 2001).

In high dimensional problems, the main concern is on the prior specification of hierarchical variance parameters since it controls the rate of shrinkage. Generally, a well defined shrinkage prior is a prior with heavy tails like Cauchy in order to allow strong effects remain large and also provides severe shrinkage for weak effects. These properties can be achieved by imposing global and local shrinkage parameters to the model. In order to have both global and local shrinkage parameters, shrinkage prior applies parameter expansion technique. Overparameterization reduces dependence among the parameters in a hierarchical model and improves MCMC converges (Liu *et al.*, 1998). Adding additional parameters can also increase flexibility of applied model. This technique was originally constructed to speed up EM and Gibbs sampler computations. However, with shrinkage priors, the aim is to control the rate of shrinkage through these parameters.

In general, shrinkage priors can be represented as scale mixtures of normal

distributions as

$$\beta_i \sim N(0, \lambda_i \tau)$$

In this representation, there are two hyperpaameters :

- Global Shrinkage Parameter, τ:
  Global parameters are shared scale parameters that try to estimate overall sparsity level. This dates back to Stein (1956). These parameters reveal the presence of sparsity in the model. Therefore, global shrinkage parameters are of fundamental importance in high dimensional inference.

- Local Shrinkage Parameter, λ_i:
  Local parameters shrink locally the nonzero parameters of β. In addition, the key role of local parameters is to reduce the gravity toward zero on strong effects exercised by global parameters.

Different local-global shrinkage priors can be found in the literature, but we discuss the shrinkage behavior of some of those priors.

3.3.1 Double-Exponential Prior

One of the most common used shrinkage priors is to specify double-exponential or Laplace prior on λ_i (see, e.g., Figueiredo, 2003; Bae & Mallick, 2004; Hans, 2009). Popularity of Laplace-like priors is due to their connection with Lasso penalization method. The lasso estimate can be interpreted as the mode of posterior of β_is with independent and identical Laplace priors. This is also known as Bayesian Lasso that is represented as

$$\lambda_i \sim \exp(-\eta^2/2), \quad \tau \propto 1/\tau \tag{3.4}$$

where $\tau = \sigma^2$, the variance of error term in (3.1).

Although the Bayesian Lasso is a Bayesian representation of the Lasso, its estimate is a compromise between the Lasso and ridge regression estimates; its path moves like Lasso but is smooth like ridge regression. Moreover, speed of shrinking β_is toward zero with Bayesian lasso is between ridge and Lasso.

Hyperparameter η can be estimated by empirical Bayes strategy or placed a hyperprior on it. Park & Casella (2008) suggest to give gamma prior of the form of

$$p(\eta^2/2) = \frac{r^s}{\Gamma(s)}(\eta^2)^{s-1}\exp(-r\eta^2), \quad r > 0, s > 0, \tag{3.5}$$

to $\eta^2/2$. This choice of prior allows easy extension of Gibbs sampler of (3.4) because η^2 can simply join the other parameters without changing their full conditional distributions.

3.3.2 Generalized Double Pareto, GDP

Generalized double Pareto density is a modified version of generalized Pareto in order to be appropriate for $p \gg n$ problems. Armagan *et al.* (2013) proposed this distribution by reflecting the positive part of generalized Pareto around origin as

$$f(\beta_i|\xi,\alpha) = \frac{1}{2\xi}\left(1+\frac{|\beta_i|}{\alpha\xi}\right)^{(1+\alpha)} \qquad \xi > 0, \alpha > 0.$$

The parameter ξ is a scale parameter that controls the dispersion and α is a shape parameter that controls the tail heaviness. The GDP has Cauchy-like tails when $\alpha = 1$. This avoids over-shrinkage on marker effects away from the origin. Figure 3.1 compares generalized double Pareto in the case that $\xi = \alpha = 1$ with Laplace and Cauchy.

Hierarchical representation of GDP as local-global shrinkage prior is

$$\lambda_i \sim \text{Exp}(\theta_i^2/2), \quad \tau \propto 1/\tau, \text{and}$$

$$\theta_i \sim \text{Gamma}(\alpha, \eta)$$

where $\xi = \dfrac{\tau^{1/2}\eta}{\alpha}$, $\alpha > 0$, $\eta > 0$ and $\tau = \sigma^2$, the variance of error term in (3.1). The rate of shrinkage has been affected by the choice of hyperparameters α and η. For ensuring the continuity property to avoid instability

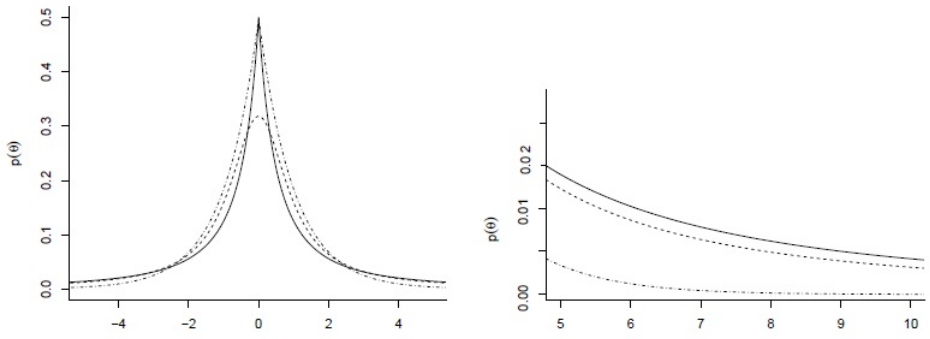


Figure 3.1: The GDP (solid line) , double-exponential prior (dash-dotted line) and standard Cauchy (dashed line).

in prediction, it is necessary and sufficient to select $\eta = \sqrt{\alpha + 1}$. With this choice, α is the only hyperparameter to specify. Picking $\alpha = 3$ induces a lighter tails than Cauchy distribution, while $\alpha = 1$ provides Cauchy-like tail prior. Letting $\alpha \to \infty$ leads to an improper prior.
Hierarchical representation of GDP makes it very similar to normal-exponential-gamma family of priors proposed by Griffin & Brown (2007). The differences is that here the mixing prior is placed on θ_i instead of θ_i^2 in the prior of Griffin and Brown. This mixing leads to simpler analytic forms for the marginals. Simple data augmentation Gibbs sampler of GDP can be obtained via the scale mixture of normals representation.

3.3.3 Horseshoe Prior

Horseshoe prior is a global-local shrinkage prior introduced by Carvalho *et al.* (2010) as

$$\lambda_i^{1/2} \sim C^+(0,1), \quad \tau \sim C^+(0,1), \quad \sigma \propto 1/\sigma,$$

where $C^+(0,1)$ is a half-Cauchy distribution. The horseshoe prior $\pi(\beta_i|\tau)$ does not have closed-form representation but it behaves like $\log(1 + 2/\beta_i^2)$. The main difference of horseshoe prior with aforementioned shrinkage priors is that the global shrinkage parameter is not the same as the variance of error term, σ^2. Separating τ and σ^2 provides more appealing features for this prior. As it is shown in Figure 3.2, flat Cauchy-like tail of horseshoe prior avoids

over shrinkage of large β_is or strong effects, while its infinity tall spike at the origin shrinks severely the low effects towards zero. The figure shows

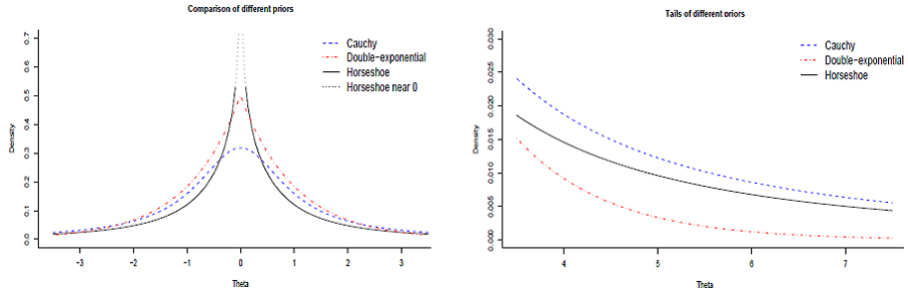


Figure 3.2: The horseshoe prior (solid line), double-exponential prior (dash-dotted line) and standard Cauchy (dashed line).

behavior of horseshoe priors with respect to two commonly used shrinkage priors, double-exponential and the Cauchy priors. The double-exponential prior causes severe shrinkage on low effects and the Cauchy reduces imposed bias on larger β_is since it has heavier tail than two the others.

3.3.4 Shrinkage Coefficient

Shrinkage coefficient, κ_i, is a random parameter that its behavior provides an understanding about the way of shrinkage. On the other words, it is the amount of weight that the posterior mean of β_i given y places on zero. Under local shrinkage prior $\beta_i \sim N(0, \lambda_i)$, the posterior mean is

$$E(\beta_i \mid y_i, \lambda_i) = (1 - \kappa_i) y_i,$$

where $\kappa_i = \dfrac{1}{1+\lambda_i}$. For $\kappa_i = 0$ there is no shrinkage and for $\kappa_i = 1$ we have total shrinkage toward origin. This is a motivation to compare different shrinkage prior through behavior of κ_i in a *priori*. Table 3.1 lists density of κ_i associated with prior of λ_i for aforementioned shrinkage priors and Cauchy prior. Presented priors are obtained up to the constant and for GDP, $\xi = \alpha = 1$ is considered. Figure 3.3 also shows the shape of these priors.

Table 3.1: Priors of κ_i associated with some shrinkage prior where $\kappa_i^* = \kappa_i/2(1-\kappa_i)$ and Erfc(.) denotes the complementary function. In addition, for GDP $\xi = \alpha = 1$.

prior for β_i	Density for κ_i
Cauchy	$\frac{\sqrt{\kappa_i^*}}{\kappa_i^{3/2}(1-\kappa_i)} e^{-\kappa_i^*}$
Double-exponential	$\kappa_i^{-2} e^{-1/(2\kappa_i)}$
GDP	$\frac{\sqrt{\kappa_i^* \pi}\text{Erfc}\left[\sqrt{\kappa_i^*}\right] e^{\kappa_i^*}}{2\kappa_i^3} - \frac{\kappa_i^*}{\kappa_i^2}$
Horseshoe	$\kappa_i^{-1/2}(1-\kappa_i)^{-1/2}$

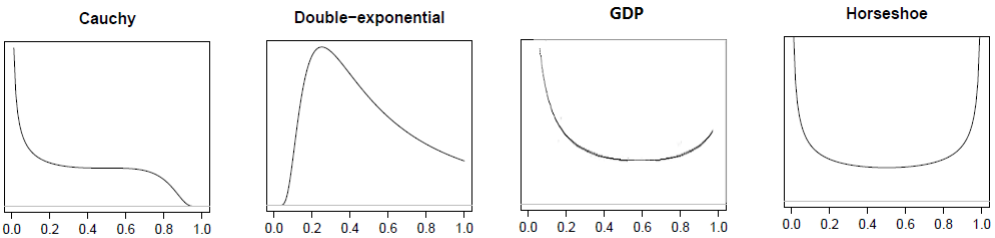


Figure 3.3: Densities of κ_i for Cauchy, Double-exponential, generalized double Pareto (GDP), and horseshoe priors.

The double-exponential prior tends to a fixed constant near $\kappa_i = 1$. This limits the ability of the prior to squelch noise components back to zero. In addition, the density vanishes entirely near $\kappa_i = 0$; it is not a good feature for shrinkage priors since no shrinkage for large effects is desired.

Horseshoe prior implies $\kappa_i \sim \text{Beta}\left(\frac{1}{2}, \frac{1}{2}\right)$. Since $\text{Beta}\left(\frac{1}{2}, \frac{1}{2}\right)$ is symmetric and unbounded at boundary points, no shrinkage near zero and total shrinkage near one is expected.

The GDP behaves similar to horseshoe near zero but its behavior is not unbounded like horseshoe and not a fixed constant like double-exponential in the neighborhood of one; i.e. it behaves between these two cases. As it is represented in Armagan *et al.* (2013), the general density of κ_i of the GDP is a

function of hyper-parameters α and η. These two hyper-parameters influence differently on κ_i near one. Increasing α places more and more density near one while increasing η places less and density near one. Therefore, these hyper-parameters should be chosen carefully.

3.3.5 Tail Robustness

Tail robustness is a property of an estimator on its behavior in situations where y is very different from the prior mean. Observing the behavior of the posterior expectation of β for large y provides insight into investigating this property. However it is not a new concept, the following theory by Carvalho *et al.* (2010) characterizes this property by relaxing boundedness condition on $\pi(\beta)$.

THEOREM: Let $p(\mid y-\beta \mid)$ be the likelihood, and suppose that $\pi(\beta)$ is a zero mean scale mixture of normals, $\beta \mid \lambda \sim N(0,\lambda)$, with $\lambda^{1/2}$ having proper prior $\pi(\lambda^{1/2})$. Assume further that the likelihood and $\pi(\beta)$ are such that the marginal density $m(y)$ is finite for all y. Define the following pseudo-densities, which may be improper,

$$m^*(y)=\int_{\mathbb{R}} p(\mid y-\beta \mid)\pi^*(\beta)d\beta, \quad \pi^*(\beta)=\int_{\mathbb{R}^+}\pi(\beta \mid \lambda)\pi^*(\lambda^{1/2})d\lambda^{1/2},$$

$$\pi^*(\lambda^{1/2})=\lambda\pi(\lambda^{1/2}).$$

Then

$$E(\beta \mid y)=\frac{m^*(y)}{m(y)}\frac{d}{dy}\log m^*(y)=\frac{1}{m(y)}\frac{d}{dy}m^*(y).$$

□

In the case that $p(\mid y-\beta \mid)$ is a normal likelihood, then $E(\beta \mid y)$ reduces to

$$E(\beta \mid y)=y+\frac{d}{dy}\log m(y), \tag{3.6}$$

(Masoeliez, 1975; Polson, 1991; Pericchi & Smith, 1992).

To achieve tail robustness, the second term in (3.6) needs to converge to zero for large $\mid y \mid$. In the case that variance of observations is one and

$\pi(\lambda) \sim \lambda^{s-1} e^{-\zeta\lambda} L(\lambda)$ such that $L(t\lambda)/L(\lambda) \to 1$ when $\lambda \to \infty$ for any $t > 0$, we have

$$\frac{d}{dy} \log m(y) \sim \frac{2^r s - 1}{y} - \sqrt{2\zeta}, \tag{3.7}$$

where $r = 1$ if $\zeta > 0$ and $r = 0$ otherwise (Polson & Scott, 2010). Equations (3.6) and (3.7) lead to

$$\lim_{y \to \infty} \big(y - \mathrm{E}(\beta \mid y)\big) = \sqrt{2\zeta},$$

i.e., any scale mixture that places exponential prior or lighter tails on $\pi(\lambda)$, always shrinks all observations to zero, no matter how far they are from zero. But by placing priors with heavier tails on λ like Cauchy in the horseshoe, the second term in (3.6) converges to zero for large observations (Carvalho *et al.*, 2010).

3.4 Mixture Prior

Another prior specification, which is widely applied in high dimensional settings, is mixture prior:

$$\beta_i \sim \sum_j \pi_j D(0, \sigma_j^2),$$

where D denotes a distribution with mean zero and variance σ_j^2. Here π_j is the prior probability of D with σ_j^2 when $\sum_j \pi_j = 1$. Although, mixture priors are comprehensible and adapted for high dimensional problems; they face some challenges in applications. For instance, label switching is a well known problem that arises with mixture priors (see e.g., Diebolt & Robert, 1994; Redner & Walker, 1984). To overcome this problem, different constraints, known as identifiability constraints, were suggested to be imposed to the model. In the case of a mixture of two normals with representation

$$\beta_i \sim (1 - \pi) N(0, \sigma_0^2) + \pi N(0, \sigma_1^2),$$

a common constraint is to restrict σ_0^2 such that $\sigma_0^2 < \sigma_1^2$. Another approach is to reparameterize the prior so that the variance of one of the components

is a scaled version of the variance of the other component as $\sigma_0^2 = \tau^{-1}\sigma_1^2$ where $\tau > 1$.

The complexity of MCMC implementation for a large p is another challenge for dealing with mixture priors. To overcome this problem, let's put the constraint $|\beta_i| > \omega_i$ on the size of β_is in order to exclude the markers that do not have strong impact on complex traits. This can be achieved if $N(0, \sigma_0^2) > N(0, \sigma_1^2)$ on intervals $(-\omega_i, \omega_i)$ where $N(0, \sigma_0^2)$ is a prior on the set of small effects and $N(0, \sigma_1^2)$ is a prior on the set of large effects. Hence, this constraint leads to

$$\log(\sigma_1/\sigma_0)/(\sigma_0^{-1} - \sigma_1^{-1}) = \omega_i^2.$$

Based on such a constraint, σ_0 cannot reach zero. This motivates to apply a mixture of a continuous distribution and a point mass at zero as

$$\beta_i \sim \pi D + (1 - \pi)\delta_0. \tag{3.8}$$

While the point mass mixture prior (3.8) is widely used in variable selection and model choice problems, it is typical to introduce a vector of latent variables

$$\gamma = (\gamma_1, ..., \gamma_p)^T$$

in the model. The γ so-called inclusion indicators is a *zero* and *one* variable corresponding to small or large β_i respectively. This modification brings convenient having a singular prior instead of mixture prior given γ. Hence, the modified (3.8) with D as a normal distribution is

$$\beta_i \mid \gamma \sim \gamma_i N(0, \sigma^2) + (1 - \gamma_i)\delta_0. \tag{3.9}$$

From (3.9), it is clear that γ has critical influence on the analysis since it keeps a subset of predictors in the model. A common prior for the inclusion indicators is $p(\gamma) = \pi^{p_\gamma}(1-\pi)^{p-p_\gamma}$ (George & McCulloch, 1993, 1997; George & Foster, 2000) where π is inclusion probability for each predictor and p_γ is the number of nonzero predictors in the model given γ. The choice of two hyperparameters π and σ^2 impacts on inference; therefore, dealing appropriately with these unknown parameters is crucial.

3.4.1 Hyper-prior on Parameters of $\boldsymbol{\beta}_\gamma$

To place a hyper-prior on parameters of $\boldsymbol{\beta}$, let rewrite the prior as

$$\boldsymbol{\beta}_\gamma \mid \gamma \sim N(0, \Sigma_\gamma) \tag{3.10}$$

where $\boldsymbol{\beta}_\gamma$ is a set of nonzero regression coefficients given γ. If γ is given, variance of $\boldsymbol{\beta}_\gamma$ is the only parameter that should be specified. This hyper-parameter reflects the size of sparsity of the model and acts like a penalization parameter. Although we might desire to control the strength of shrinkage by placing a hierarchical prior, here we discuss about g-prior that is widely adopted for point mass mixture priors. The g-prior is one of the most popular priors on the scale parameter of the normal which is introduced by Zellner (1986) as

$$\Sigma_\gamma = \sigma^2 g(\mathbf{X}_\gamma^T \mathbf{X}_\gamma)^{-1}$$

where $p(\sigma^2) \propto 1/\sigma^2$. Here, the covariance matrix of $\boldsymbol{\beta}_\gamma$ is a scalar multiple g of the Fisher information matrix, which depends on the observed data through the design matrix $\mathbf{X}$. This particular prior has been widely adopted in the context of Bayesian variable selection since availability of the closed form of all marginal likelihoods brings computational advantages. It also has simple interpretation since it can be derived from the idea of a likelihood for a pseudo-data set with the same design matrix $\mathbf{X}$ as the observed sample (see, Zellner, 1986; George & Foster, 2000; Smith & Kohn, 1996; Fernandez *et al.*, 2001).

By the choice of g-prior g is the only hyper-parameter that needs to be specified. The hyper-parameter g acts like a penalization parameter and effectively influences on the inference. Therefore, different choices of g are recommended in literature.

- Kass & Raftery (1995) chose $g = n$ with the belief that the amount of information about the parameter should be equivalent to the amount of information in one observation which is defined through Fisher information. The result of this choice is very close to BIC criterion.

- Foster & George (1994) recommended the choice of p_γ^2 for g based on the Risk Inflation Criterion (RIC).

- Fernandez *et al.* (2001) made a bridge between BIC and RIC by the choice of $g = \max(n, p_\gamma^2)$. They named this prior benchmark prior.

- George & Foster (2000) and Cui & George (2008) estimated g via empirical Bayes method.

Those such fixed values for g may not bring ability to control the strength of shrinkage. Instead, it is more natural to consider uncertainty of this parameter by placing a prior on g.

- Zellner & Siow (1980) considered Inverse-Gamma distribution as hyperprior on g. Despite the g-prior, Zellner-Siow prior does not provide closed form of the marginal likelihoods.

- Liang *et al.* (2008) placed prior on shrinkage factor of g-prior as

$$\frac{g}{1+g} \sim \text{beta}(1, \frac{a}{2} - 1)$$

  where $a \in (2, \infty)$. For $a = 4$, the prior on the shrinkage factor is a uniform prior. For any values greater than 4 the prior places more mass around zero. Conversely, for $a \in (2, 4)$, prior of shrinkage factor concentrates around one.

In some problems like the GWAS, $p_\gamma \geq n$ for many submodels. In such cases, the matrix $X_\gamma^T X_\gamma$ is not invertible; therefore, criteria such as AIC, BIC and RIC will be unavailable for all submodels. To avoid this problem, Maruyama & George (2011) introduce generalized g-prior by decomposing X_γ via singular value decomposition technique.

The g-prior and generalized g-prior assume that regression coefficients or marker effects are not independent. While in the GWAS β_is are reflecting causal effect of $\mathbf{x}_i$s on Y, it might be better not to assume the same correlation structure of $\mathbf{x}_i$s for β_is. Thus, hereafter, we assume that β_is are independent.

3.4.2 Hyper-priors on Inclusion Probability

Inclusion Probability can be fixed with any values between $(0, 1)$, but placing a prior on π can provide more flexibility. Ley & Steel (2007) has shown that the hierarchical prior on π outperforms the prior with fixed inclusion probabilities. The common choice of prior for π is a beta distribution, Beta(a, b) (see e.g., Cui & George, 2008; Scott & Berger, 2010) which induces

$$p(\gamma) = \text{Beta}\left(p_\gamma + a, p + p_\gamma + b\right).$$

By the choice of $a = b = 1$ the prior will be $U(0, 1)$ where U denotes the uniform distribution. However, uniform distribution has been applied in some applications but it may not be appropriate for high dimensional problems. Considering a uniform prior for π corresponds to have large number of nonzero β in a *priori*. Hence, it may not be a good choice for very sparse problems. Instead, Guan & Stephens (2011) considered a uniform prior for $\log(\pi)$ as

$$\log(\pi) \sim U(\log(1/p), \log(M/p)).$$

This uniform prior is defined through prior guess of range of π in $(1/p, M/p)$. Although the bounds are a function of p to span the order of magnitude for larger p, the choice of M is not well defined.

3.5 Simulation Study

In this section, we compare out-of-sample predictive performance of Ridge Regression (RR) and the Lasso (L) as penalization methods, double-exponential prior by Park & Casella (2008) that we call it Bayesian Lasso (BL), generalized double Pareto (GDP) and horseshoe (HS) as Bayesian approaches. We also made comparison with Student's t prior that can be expressed as a mixture of normal distributions $\beta_i \sim N(0, \lambda_i)$ with the mixing scaled inverse-gamma distribution,

$$\lambda_i \sim \text{Inv-gamma}(\nu, s^2). \tag{3.11}$$

This prior specification is so called Bayesian ridge (BR). For this experiment, we consider the following scenarios for $n = 100$.

Model-1: Ten regression coefficients are 3 and the rest are zero when $p = 100$.
Model-2: Ten regression coefficients are 3 and the rest are zero when $p = 200$.
Model-3: Fifty nonzero regression coefficients, 25 generated from $U(-3, -2)$ and 25 from $U(2, 3)$, when $p = 1000$.
Model-4: All regression coefficient generated from $U(-1, 1)$ when $p = 100$.
Model-5: All regression coefficient generated from $U(-1, 1)$ when $p = 200$.
Model-6: All regression coefficient generated from $U(-1, 1)$ when $p = 1000$.

The three first models are referred to the sparse problems while three last models are dense problems motivated by genome-wide association studies.

In this study, we aim to generate a set of data similar to SNP-data. Since in GWAS markers are in linkage disequilibrium which vanishes by physical distance in genome, predictors $\mathbf{x}_1, ..., \mathbf{x}_p$ were first simulated according to a central multivariate normal distribution such that covariance between $\mathbf{x}_i$ and $\mathbf{x}_j$ is $0.5^{|i-j|}$. Additionally, markers in SNP-data set gets $0, 1$ and 2 when 1 is two times more probable than 0 and 2 to occurs. Thus, $\mathbf{x}_i$ is trichotomized as $0, 2$ and 1 if it is smaller than $\Phi^{-1}\left(\frac{1}{4}\right)$, larger than $\Phi^{-1}\left(\frac{3}{4}\right)$ or in between. The response $\mathbf{y}$ was finally simulated from a linear regression model for each scenarios when the noise term is normally distributed centered at zero with variance one. After simulating data, $\mathbf{y}$ and $\mathbf{x}$ is centered and $\mathbf{y}$ are standardized to have unit variance.

To implement MCMC for Bayesian methods, 6000 first samples out of 12000 were considered as burn-in samples. Hence, the inference has been based on 6000 remained samples. For two penalized methods, tuning parameters have been estimated from 10-fold cross validated.

To illustrate this example we used package `MASS` and `Lars` for Ridge regression and the Lasso respectively. For Horseshoe we modified the code in package `monomvn` for our model.

Table 3.2: First rows of sparse models present average of 50 out-of-sample MSPEs for RR, L, BL, BR, GDP and HS and their standard deviation based on bootstrap samples in subscript. Second rows present average of 50 correlations between prediction and observed values in validation sets.

Model		RR	L	BR	BL	GDP	HS
Model-1	MSPE	$0.107_{(.008)}$	$0.042_{(.001)}$	$0.102_{(.016)}$	$0.0516_{(.008)}$	$0.093_{(.010)}$	$0.041_{(.004)}$
	Cor	0.9204	0.990	0.960	0.983	0.978	0.991
Model-2	MSPE	$0.240_{(.017)}$	$0.049_{(.005)}$,	$0.191_{(.018)}$	$0.107_{(.010)}$	$0.154_{(.013)}$	$0.046_{(.006)}$
	Cor	0.653	0.988	0.912	0.965	0.932	0.989
Model-3	MSPE	$0.893_{(.018)}$	$0.841_{(.026)}$	$0.943_{(.030)}$	$0.831_{(.033)}$	$0.912_{(.006)}$	$0.891_{(.031)}$
	Cor	0.0827	0.374	0.170	0.352	0.224	0.244

For each model, we simulated 50 data set. We evaluated the out-of-sample predictive performance of each model, by randomly selected 80% of samples as training set and the rest as validation set. Table 3.2 and 3.3 represents average of mean square prediction error, MSPE, of 50 simulated data for each model. The index number shows the average standard error of each MSPE obtained by averaging 200 bootstrap samples of 50 standard error of each model. We also represented average of correlation between predicted and observed values in validation sets. This gives better sight to accuracy of performance for models with different number of predictors.

Table 3.2 shows that HS outperforms its competitors for sparse models with $p = n$ and $p = 2n$. However, the lasso performs very close to HS; the performance of BL is not as good as the lasso since it is compromises between ridge and the lasso. For *Model-3* with $p = 10n$, all approaches show very poor performance and their correlations reveal that any inference based on these results may not be reliable.

Table 3.3: First rows of dense models present average of 50 out-of-sample MSPEs for RR, L, BL, BR, GDP and HS and their standard deviation based on bootstrap samples in subscript. Second rows present average of 50 correlations between prediction and observed values in validation sets.

Model		RR	L	BR	BL	GDP	HS
Model-4	MSPE	$0.252_{(.014)}$,	$0.354_{(.024)}$	$0.205_{(.015)}$	$0.295_{(.026)}$	$.308_{(.012)}$	$0.526_{(.027)}$
	Cor	0.870	0.780	0.868	0.830	0.824	0.674
Model-5	MSPE	$0.496_{(.022)}$,	$0.735_{(.032)}$,	$0.466_{(.032)}$	$0.525_{(.043)}$	$0.487_{(.014)}$	$0.734_{(.030)}$
	Cor	0.696	0.519	0.701	0.643	0.682	0.486
Model-6	MSPE	$0.827_{(.019)}$	$0.971_{(.039)}$	$0.832_{(.020)}$	$0.890_{(.030)}$	$0.885_{(.007)}$	$0.979_{(.041)}$
	Cor	0.358	0.152	0.355	0.262	0.339	0.137

Table 3.3 represents the performance of the methods for dense models. As it is shown, BR has better performance in comparison with other competitors, even better than RR. However, MSPE increases by increasing p to $2n$ in *model-5* for all approaches, performance of GDP turns out to be better than RR and very close to BR. It also has smaller standard deviation than BR. Therefore, GDP can be an alternative to BR for dense problems with $p > n$. For *Model-6*, we have the same problem as *Model-3*, large MSPE and small correlation.

If we make a comparison between the sparse problems and dense problems, it is clear that for sparse problems we have better performance.

As we realized from our analysis presented in chapter 6, and previous studies on SNP-data, genome-wide problems are more similar to the dense models. Thus, we find BR and GDP as a good choice for our purpose.

Chapter 4

Bayesian Compressed Regression

In the previous chapter, we have explored Bayesian approaches based on shrinkage priors and mixture priors. Between those two categories, a shrinkage prior might be more appropriate choice for SNP-data, while there is a general agreement that most complex traits are affected by a large number of small-effect markers. Although shrinkage priors are developed for high dimensional problems, they might not provide so accurate results when we face ultrahigh dimensional problems. Shrinkage priors can be safely applied for problems with 2 or 3 times more predictors than observations. However, the true and safe upper limit is specific for each problem due to degree of correlation (co-linearity) among the predictors. Hence, in the first step, reducing the dimensionality of the set of SNPs is required.

In this chapter, we focus on random projection method that is a dimensional reduction technique. Random projection compresses the data in low dimension space in such a way that we can learn about the complex trait from compressed data with little loss of information. The aim of this chapter is to understand whether random projection can be appropriate for SNP-data. Hence, we first review the main concepts of random projection. In second section, we modified Bayesian compress regression by Guhaniyogi & Dunson (2013) for the problems with related samples. The section three presents

prediction model when we have related samples. In the section four, sensitivity of analysis based on random projection is discussed. The last section is devoted to illustrate a simulation study to evaluate out-of-sample predictive performance of compressed regression and Bayesian shrinkage prior with presence of random effects.

4.1 Dimensional Reduction

A widely used approach to dealing with large p is to first reduce the dimension with a dimensionality reduction techniques. Principal component analysis, PCA, is an extremely important tool for this aim which has found use in many experimental and theoretical studies. The PCA finds an m-dimensional subspace of $\mathbb{R}^p$ which captures as much of the variation in the data set as possible(Maruyama & George, 2011; Paul *et al.*, 2008; Li *et al.*, 2011).

Although the PCA is based on linear mapping, its computational complexity precludes its use in truly large-scale applications. The computational complexity of PCA is $O(p^2n) + O(p)$ which makes it inefficient for a problem like the GWAS. Although, computing singular value decomposition, SVD, is somewhat more efficient, it is still expensive for large p.

An alternative to the PCA is random projection, RP, which is also based on linear mapping. Performing random projection requires only a matrix multiplication and takes $O(npm)$. Due to the low computational cost of the RP, it gets some attention in literature during the last two decades. It reduces dimension of the parameter space down to $O(\log p)$ with linear computational cost in the dimension . Therefore, the RP has been found computationally efficient and a sufficiently accurate method for dimensionality reduction of high dimensional settings. The main idea of the RP is to reduce the dimensionality by projecting the data into a lower dimensional subspace formed by a set of random vector. Let $X \in \mathbb{R}^{n\times p}$ be our n points in p dimensions. To reduce the p down to m that is much smaller than p, the random projection

technique multiplies X by a random matrix $\Phi \in \mathbb{R}^{p\times m}$, as

$$\tilde{X} = \frac{1}{\sqrt{m}} X\Phi, \qquad \tilde{X} \in \mathbb{R}^{n\times m}. \tag{4.1}$$

Here, the entries of Φ, ϕ_{ij}s, should be generated independently from an identical symmetric α-stable distribution.

A random variable ϕ_{ij} is called symmetric α-stable random variable if its characteristic function is

$$E\left[\exp(\sqrt{-1}\phi_{ij}t)\right] = \exp(-d \mid t \mid^{\alpha})$$

where $d > 0$ is scale parameter. In the case that $\alpha = 2$ and $\alpha = 1$, ϕ_{ij} is a normal and Cauchy random variable respectively. The choice of α brings the properties of L_α norm distance. For instance, if $\alpha = 2$, the L_2 distance between the rows of original data X is preserved in the projected matrix $\tilde{X}$.

One of the main issues in random projection is to determine m. To this end, let $\mathbf{x}_i$ denotes ith row of matrix X and $\tilde{\mathbf{x}}_i$ denotes ith row of $\tilde{X}$. For convenience, we focus on the leading two rows, $\mathbf{x}_1$ and $\mathbf{x}_2$ in X, and the leading two rows, $\tilde{\mathbf{x}}_1$ and $\tilde{\mathbf{x}}_2$, in $\tilde{X}$ such that

$$s_1 = \parallel \mathbf{x}_1 \parallel_\alpha = \sum_{j=1}^{p} \mid x_{1j} \mid^{\alpha}, \quad s_2 = \parallel \mathbf{x}_2 \parallel_\alpha = \sum_{j=1}^{p} \mid x_{2j} \mid^{\alpha},$$

$$d_\alpha = \parallel \mathbf{x}_1 - \mathbf{x}_2 \parallel_\alpha = \sum_{j=1}^{p} \mid x_{1j} - x_{2j} \mid^{\alpha} .$$

A typical method to choose m is to bound

$$P\left(\mid \hat{d}_\alpha - d_\alpha \mid > \varepsilon d_\alpha\right)$$

where $\hat{d}_\alpha$ is estimate of d_α and ε is a factor to control accuracy of the projection to preserve L_α distance. This central idea of random projection dates back to Johnson & Lindenstrauss (1984). It proves that in particular case $\alpha = 2$,

$$(1-\varepsilon)d_2 \le \hat{d}_2 \le (1+\varepsilon)d_2, \quad \hat{d}_2 = \parallel \tilde{\mathbf{x}}_1 - \tilde{\mathbf{x}}_2 \parallel_2$$

holds for $m \ge m_0 = O(\varepsilon^{-2}\log n)$, which is so called JL-Lemma. In other words, JL-lemma says that if we perform an orthogonal projection of n points

in a vector space $\mathbb{R}^p$ onto a selected lower-dimensional subspace, then L_2 distances between points are preserved; i.e., not distorted more than a factor of $(1 \pm \varepsilon)$, for any $0 < \varepsilon < 1$. Various JL-Lemma (see, e.g., Dasgupta & Gupta, 2003; Indyk & Motwani, 1998) have been proved for precisely determining m given some specified level of accuracy.

4.1.1 The Choice of Random Projection Matrix

A crucial point in random projection is to select a method to generate random ϕ_{ij} since its distribution can change the variances (average errors) and error tail bounds. This task is equivalent to the choice of α that must be chosen based on the data and the problem in hand. Here, we just consider common random projection when $\alpha = 2$.

Normal random projection

By the choice of $\alpha = 2$ the random projection matrix has i.i.d normal entries. This kind of random projection is the simplest random projection in terms of analysis; although it is not the simplest from generating a random number view point. For this choice of random projection matrix, we can easily show that $k \parallel \tilde{\mathbf{x}}_1 \parallel^2 / s_1 \overset{\mathrm{D}}{=} \chi^2_k$ where χ^2_k denotes chi-square distribution with k degree of freedom. Based on this result, we can learn more about concentration of $\parallel \tilde{\mathbf{x}}_i \parallel^2$ around its expectation,$\parallel \mathbf{x}_i \parallel^2$. Vempala (2004) readily shows

$$E(\parallel \tilde{\mathbf{x}}_1 \parallel_2) = s_1 \quad \text{and} \quad E(\parallel \tilde{\mathbf{x}_1} - \tilde{\mathbf{x}_2} \parallel_2) = d_2$$

$$\text{var}(\parallel \tilde{\mathbf{x}}_1 \parallel_2) = \frac{2}{m} s_1^2, \quad \text{var}(\parallel \tilde{\mathbf{x}}_1 - \tilde{\mathbf{x}}_2 \parallel_2) = \frac{2}{m} d_2^2$$

that gives a hint about how to choose m; i.e. with appropriate choice of m we can expect an accurate result.
Invoking to well-know Chernoff-inequality, we have

$$P\left(\mid \hat{d}_2 - d_2 \mid \geq \varepsilon d_2\right) \leq 2 \exp\left(-\frac{m}{4}\varepsilon^2 + \frac{m}{6}\varepsilon^3\right)$$

for any $0 < \varepsilon < 1$. By applying Bonferroni union bound as

$$\frac{n^2}{2} 2 \exp\left(-\frac{m}{4}\varepsilon^2 + \frac{m}{6}\varepsilon^3\right) \leq n^{(-\zeta)}$$

then

$$m > m_0 = \frac{4 + 2\zeta}{\varepsilon^2/2 + \varepsilon^3/3} \log n \tag{4.2}$$

where $1 - n^{-\zeta}$ is the probability of holding JL-Lemma (Achlioptas, 2003). On the other words, with this lower bound for m, we control the probability of success of JL lemma by ζ while ε controls the desired accuracy in distance preservation.

Sub-Gaussian random projection

Although normal random projection is simple in theory, it is not an efficient choice for large p because of its computational complexity due to the dense nature of the projection matrix. Instead, generating a projection matrix from sub-Gaussian distribution is convenient computationally and theoretically.

A variable x is said a sub-Gaussian random variable if

$$E\left[\exp(tx)\right] \leq \exp\left(\frac{\upsilon^2 t^2}{2}\right), \forall t \in \mathbb{R} \tag{4.3}$$

for some $\upsilon > 0$. In other words, if there is a positive real number υ such that the Laplace transform of x is dominated by the Laplace transform of a Gaussian random variable with mean zero and variance υ^2, then x is a sub-Gaussian variable with parameter υ^2.

Hence, we can easily generate ϕ_{ij}s from any zero-mean bounded variance distribution. A typical choice of this kind of projection matrix is

$$\phi_{ij} = \sqrt{s} \begin{cases} 1, & \text{with probability} \quad 1/2s \\ 0, & \text{with probability} \quad 1 - 1/s \\ -1, & \text{with probability} \quad 1/2s. \end{cases} \tag{4.4}$$

where $s \geq 1$. It is obvious that this choice of random projection is computationally appealing since $1/s$ fraction of data is only projected in new space. The other words, it is s-fold speedup random projection. Achlioptas (2003) first suggested the use of (4.4) for $s = 1, 3$. He dropped the spherical condition of JL-Lemma by presenting new version of JL-Lemma. For any random

projection matrix generated from unite variance sub-Gaussian distribution with parameter ν^2, we have

$$P\left(\hat{d}_2 \leq (1+\varepsilon)d_2\right) \leq \exp\left[-\frac{m}{2}\left(\log\frac{\delta^2}{1+\varepsilon} + \frac{1+\varepsilon}{\delta^2} - 1\right)\right], \quad \varepsilon > 0 \quad (4.5)$$

where δ_ϕ^2 is an optimal value of ν^2 and $\nu^2 \leq 1+\varepsilon$. This upper bound can be obtained by using Chernoff inequality

$$P\left(\hat{d}_2 - d \geq \varepsilon d\right) \leq \frac{E\left[\exp\left(\hat{d}_2 t\right)\right]}{\exp\left[(1+\varepsilon)d_2 t\right]}.$$

Since the upper bound is a function of m, a good choice of m can ensure that the inequality (4.5) holds with high probability.

For very sparse problem, Li *et al.* (2006) suggested to generate random variable from (4.4) with $s > 3$. They have shown that under some conditions the upper bound (4.5) can be even reached for s up to $\sqrt{p}$.

4.2 Bayesian Compressed Regression

Linear regression model with presence of random effects is widely applied in genetic problems. This model is capable of correcting for several forms of confounding due to genetic relatedness such as population structure and familial relatedness. Therefore, hereafter, we consider

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim N(0, \tau^{-1}R) \quad (4.6)$$

where $\mathbf{y}$ is an n vector of quantitative trait measured on n experimental units, $\mathbf{X}$ is an $n \times p$ matrix of genotypes measured at genetic markers, $\boldsymbol{\beta}$ is a p vector of additive genetic effects and R is an $n \times n$ known diagonal matrix so called heterogeneous-residual variance. Random vector $\mathbf{u}$ is an n vector such that

$$\mathbf{u} \sim N(\mathbf{0}, \tau^{-1}K)$$

where K is an $n \times n$ known relatedness matrix calculated from pedigree. Variance covariance matrix of $\mathbf{u}$ is a function of relatedness matrix in order to consider genetic relation among experimental units.

In compressed regression, we project each row $\mathbf{x}_i \in \mathbb{R}^p$ of $\mathbf{X}$ to an m dimensional subspace through an $p \times m$ Normal random projection matrix Φ since our model is a normal linear model. After projecting data in new space, $\mathbf{y}$ is regressed on projected design matrix $\mathbf{X}\Phi$. Hence, (4.6) is modified as

$$\mathbf{y} = \mathbf{X}\Phi\tilde{\boldsymbol{\beta}} + \mathbf{u} + \boldsymbol{\epsilon} \tag{4.7}$$

where $\tilde{\boldsymbol{\beta}}$ is an m vector of coefficient in projected space. As suggested by Guhaniyogi & Dunson (2013), after generating random matrix projection, we assume that Φ is fixed. Then, we consider two different scenarios

- large p and large n
  For large enough n, we can find $m_0 < n$ such that (4.2) holds for small ε with high probability. In such a problem, a usual conjugate priors can be considered for the model, while we are not in high dimensional setting anymore. In particular, we choose following priors for parameters
  $$\tilde{\boldsymbol{\beta}} \sim N(0, \tau^{-1}\Sigma_\beta), \qquad \tau \sim \text{Gamma}(a_1, b_1). \tag{4.8}$$
  Then we obtain the posterior distribution of τ given Φ as
  $$\tau \mid \mathbf{y}, \Phi \sim \text{Gamma}(n/2 + a, b + \mathbf{y}^T(R + K + \mathbf{X}\Phi\Sigma_\beta\Phi^T\mathbf{X}^T)^{-1}\mathbf{y})$$
  and the posterior distribution of $\tilde{\boldsymbol{\beta}}$ and $\mathbf{u}$ given Φ as
  $$\tilde{\boldsymbol{\beta}} \mid \mathbf{y}, \Phi \sim t_n\left(W^{-1}\Phi^T\mathbf{X}^TA^{-1}\mathbf{y}, \quad 2\frac{b_1}{n}W^{-1}\right),$$
  $$\mathbf{u} \mid \mathbf{y}, \Phi \sim t_n\left(\left[B^{-1} + K^{-1}\right]^{-1}B^{-1}\mathbf{y}, \quad 2\frac{b}{n}\left[B^{-1} + K^{-1}\right]^{-1}\right)$$
  for the case that $a_1 \to 0$, $b_1 \to 0$. Here
  $$A = R + K, \quad B = \mathbf{X}\Phi\Sigma_{\tilde{\beta}}\Phi^T\mathbf{X}^T + R,$$
  $$W = \Phi^T\mathbf{X}^TA^{-1}\mathbf{X}\Phi + \Sigma_{\tilde{\beta}}^{-1},$$
  $$b = \mathbf{y}^T\left(R + K + \mathbf{X}\Phi\Sigma_{\tilde{\beta}}\Phi^T\mathbf{X}^T\right)^{-1}\mathbf{y}.$$

- large p and small n
  When n is small, for holding JL-Lemma for small ε with high probability, lower bound for m in (4.2) is greater than n. In this case, shrinkage prior is appropriate choice for model (4.7). For instance, we place generalized double Pareto on marker effects. Thus, prior specification on hyperparameters in (4.8) is

  $$\Sigma_{\tilde{\beta}} = \text{diag}\{\eta_j\}, \quad \eta_j \sim \text{Exp}(\lambda_j^2/2), \quad \lambda_j \sim \text{Gamma}(a_2, b_2), \quad j = 1, ..., m.$$

  We estimate the parameters of the model by sampling from their conditional posterior distributions through MCMC algorithm that is known as Gibbs sampling scheme. These conditional posterior distributions are given in the follow.

  $$\tilde{\boldsymbol{\beta}} \quad | \, . \sim N\left(W^{-1}\Phi^T\mathbf{X}^T R^{-1}(\mathbf{y}-\mathbf{u}), \tau^{-1}W^{-1}\right),$$

  $$\tau \quad | \, . \sim IG\left(n + p/2, \quad \frac{1}{2}\left[\left(\mathbf{y} - \mathbf{X}\Phi\tilde{\boldsymbol{\beta}} - \mathbf{u}\right)^T R^{-1}\left(\mathbf{y} - \mathbf{X}\Phi\tilde{\boldsymbol{\beta}} - \mathbf{u}\right) + \tilde{\boldsymbol{\beta}}^T\Sigma_{\tilde{\beta}}^{-1}\tilde{\boldsymbol{\beta}} + \mathbf{u}^T K^{-1}\mathbf{u}\right]\right),$$

  $$\mathbf{u} \quad | \, . \sim N\left(\left[R^{-1} + K^{-1}\right]^{-1} R^{-1}\left(\mathbf{y} - \mathbf{X}\Phi\tilde{\boldsymbol{\beta}}\right), \tau^{-1}\left[R^{-1} + K^{-1}\right]^{-1}\right),$$

  $$\eta_j^{-1} \mid . \sim \text{Inv-Guass}\left(\left(\frac{\lambda_j^2}{\tilde{\beta}_j^2\tau}\right)^{1/2}, \quad \lambda_j^2\right),$$

  $$\lambda_j \quad | \, . \sim \text{Gamma}\left(a_2 + 1, \quad \tau^{1/2} \mid \tilde{\beta}_j^2 \mid + b_2\right).$$

4.2.1 Prediction model

Some problems in the GWAS aim to predict the quantitative trait given new genotypes measured on new individuals. In these kinds of problems,

compress regression is more appealing since the initial marker-effects, $\boldsymbol{\beta}$, do not estimate directly with model (4.7).

To predict new observation, we need to consider that new experimental units might be related to observed samples. Indeed, we assume that random effects for the observed and future experimental units follow multivariate normal distribution

$$\begin{pmatrix}\mathbf{u}_o\\ \mathbf{u}_f\end{pmatrix} \sim N\left(\begin{pmatrix}0\\0\end{pmatrix}, \tau^{-1}\begin{pmatrix}K_{o,o} & K_{o,f}\\ K_{f,o} & K_{f,f}\end{pmatrix}\right).$$

The index o indicates the observed data and f represents the future one. Based on standard multivariate theory the conditional distribution of new random effects given the observed data is

$$\mathbf{u}_f \mid \mathbf{u}_o \sim N\left(K_{f,o}K_{o,o}^{-1}\mathbf{u}_o, \quad \tau_u^{-1}\left[K_{f,f} - K_{f,o}K_{o,o}^{-1}K_{o,f}\right]\right).$$

Invoking to the Law of total expectation and variance, one can obtain the posterior expectation and variance of predictive quantitative trait, $\mathbf{y}_f$, given $\mathbf{X}_f$ and $\mathbf{y}_o$. The posterior expectation of future data that we call it predictive model is given by

$$\begin{aligned}\mathrm{E}\left(\mathbf{y}_f \mid \mathbf{X}_f, \mathbf{y}_o, \Phi\right) &= \mathrm{E}\left(\mathrm{E}\left(\mathbf{X}_f\Phi\tilde{\boldsymbol{\beta}} + \mathbf{u}_f + \boldsymbol{\epsilon}_f \mid \mathbf{X}_f, \mathbf{y}_o, \Phi, \tilde{\boldsymbol{\beta}}, \mathbf{u}_o, \tau\right)\right)\\ &= \mathbf{X}_f\Phi\ \mathrm{E}\left(\tilde{\boldsymbol{\beta}} \mid \mathbf{y}_o, \Phi\right) + K_{f,o}K_{o,o}^{-1}\mathrm{E}\left(\mathbf{u}_o \mid \mathbf{y}_o, \Phi\right)\end{aligned}$$

and we write it simply as

$$\hat{\mathbf{y}_f} = \mathbf{x}_f\Phi\hat{\tilde{\beta}}_o + K_{f,o}K_{o,o}^{-1}\hat{\mathbf{u}}_o. \tag{4.9}$$

The posterior variance of prediction is also obtained as

$$\begin{aligned}\mathrm{Var}\Big(\mathbf{y}_f \mid \mathbf{X}_f, \mathbf{y}_o\Phi\Big) =& \mathbf{X}_f\Phi\mathrm{Var}\Big(\tilde{\boldsymbol{\beta}} \mid \mathbf{y}_o, \Phi\Big)\Phi^T\mathbf{X}_f^T\\ &+ K_{f,o}K_{o,o}^{-1}\ \mathrm{Var}\Big(\mathbf{u}_o \mid \mathbf{y}_o, \Phi\Big)K_{o,o}^{-1}K_{o,f}\\ &+ \Big(K_{f,f} - K_{f,o}K_{o,o}^{-1}K_{o,f}\Big)\mathrm{E}\Big(\tau_u^{-1} \mid \mathbf{y}_o, \Phi\Big)\\ &+ R_{f,f}\mathrm{E}\Big(\tau^{-1} \mid \mathbf{y}_o, \Phi\Big).\end{aligned}$$

4.2.2 Sensitivity of Inference to the choice of m

However random projection is highly efficient from computational point of view and it has shown noticeable result in high dimensional data analysis in different applications, the main drawback of random projection is its unstable result. Different random projections may lead to different inference. This problem mainly arises due to the choice of m, the dimension of projecting space. Although, many studies have been done to obtain lower bound for m, such as (4.2), it is still an open question how to choose m for a random projection in order to get an stable result.

The instability of the inference in Bayesian compressed regression can be seen as uncertainty about the model due to the ambiguity in the choice of (m, Φ). A complete Bayesian solution to this problem involves averaging over all possible models under investigated problem. Let consider M_l, $l = 1, 2, .., s$, represent lth model corresponding to Φ_l. If we are interested on prediction of future observation, y_f, the posterior distribution of y_f given $\mathbf{x}_f$ and observed data $D = (\mathbf{y}_o, \mathbf{x}_o)$ is

$$P(y_f) = \sum_{l=1}^{s} P(y_f \mid \mathbf{x}_f, M_l, D) P(M_l \mid D).$$

This is an average of the posterior distribution of y_f under each model weighted by corresponding posterior model probability which is

$$P(M_l|D) = \frac{P(D|M_l)P(M_l)}{\sum_{k=1}^{s} P(D|M_k)P(M_k)}$$

where $P(D \mid M_l)$ is marginal likelihood under model M_l and

$$P(D \mid M_l) = \int P(D \mid \theta_l, M_l) P(\theta_l \mid M_l) d\theta_l. \tag{4.10}$$

In equation (4.10), $\theta_l = (\tilde{\boldsymbol{\beta}}_l, \mathbf{u}_l, \tau_l)$, $P(D \mid \theta_l, M_l)$ is likelihood and $P(M_l)$ is the prior probability that M_l is the true model. This equation for our model is obtained as

$$\begin{aligned} P(D \mid M_l) &= \int P(D \mid M_l, \tilde{\boldsymbol{\beta}}_l, \mathbf{u}_l, \tau_l) \pi(\tilde{\boldsymbol{\beta}}_l, \mathbf{u}_l, \tau_l) d\tilde{\boldsymbol{\beta}}_l d\mathbf{u}_l d\tau_l \\ &= \frac{\Gamma\left(\frac{n}{2}\right) (\pi)^{-n/2}}{\mid R + K + \mathbf{X}\Phi\Sigma_\beta\Phi^T\mathbf{X}^T \mid^{1/2} b^{n/2}}. \end{aligned}$$

Like most of the problems, it is not practical averaging over all possible models. Guhaniyogi & Dunson (2013) suggested $\left[\lceil 2\log(p)\rceil, \min(n,p)\right]$ as a window for possible size of new space, m.

4.3 Simulation Study

In this section, we compare prediction performance of (4.7) for $n = 100$ by illustrating a simulation study; although random projection is more appropriate for large data. Let consider two different number of predictors $p = 1000$ and 2000 for two different scenarios for the size of β_is as
Model 1 : Only 20 of regression coefficients are 3 and all others are zero,
Model 2 : All regression coefficients are generated from $U(-1, 1)$.

The first model is referred to a sparse problem while the second one is a dense model which is motivated by SNP-data.

In order to simulate data similar to the real-SNP data, predictors $\mathbf{x}_1, ..., \mathbf{x}_p$ were simulated as predictors in section 3.5. Then we generated n random effects from a multivariate normal distribution with zero mean and covariance matrix K. Relatedness matrix, K, should be defined by pedigree that shows how the samples are related through $\mathbf{x}_j$s, $j = 1, ..., n$. Since for simulated data we did not have pedigree, we considered an special case that $K = \mathbf{X}\mathbf{X}^T/p$. The response $\mathbf{y}$ was finally calculated from linear mixed model for each scenarios when the noise term is normally distributed with heterogeneous-residual variance. The inverse of heterogeneous-residual variances are in $(0.7, 0.99)$ that is generated randomly.

In our experiments, $\mathbf{y}$ and $\mathbf{x}$ is centered and $\mathbf{y}$ is standardized to have unit variance. For each model, we first investigate out of sample prediction performance with three different prior specification, Bayesian lasso (BL), Bayesian Ridge (BR), and generalized double Pareto (GDP). We then projected the data to a lower dimensional space with $m = 300$. This time we implement the MCMC with BL, BR and GDP for projected data that we denoted by CBR, CBL, CGDP respectively.

Table 4.1: Rows of each model present average of 20 out-of-sample MSPEs for BR, CBR, BL, CBL, CGDP, GDP with their standard deviation based on bootstrap samples in subscript.

Model	p	BR	CBR	BL	CBL	GDP	CGDP
Model-1	1000	$0.843_{(.029)}$	$0.936_{(.035)}$	$0.635_{(.025)}$	$0.647_{(.034)}$	$0.795_{(.009)}$	$0.967_{(.018)}$
	2000	$0.901_{(.029)}$	$0.915_{(.0515)}$	$0.868_{(.050)}$	$1.003_{(.058)}$	$0.89_{(.012)}$	$1.023_{(.034)}$
Model-2	1000	$0.880_{(.011)}$	$0.899_{(.028)}$	$0.915_{(.02)}$	$1.008_{(.034)}$	$0.894_{(.009)}$	$1.030_{(.016)}$
	2000	$0.902_{(.019)}$	$0.981_{(.036)}$	$0.929_{(.021)}$	$0.964_{(.041)}$	$0.917_{(.011)}$	$1.170_{(.023)}$

For each experiment, to evaluate the out of sample performance of each model %80 of samples selected as training set and the rest considered as validation set. Table 4.1 represents average of MSPEs of 20 simulated data. The index numbers show the average of standard errors of MSPEs which obtained by averaging 100 bootstrap samples of 20 standard errors of each model.

As it is represented in table 4.1, the compress regression for all models increases the MSPEs respect to non-projected data so as to increase computational efficiency. This result reveals that Bayesian compressed regression is a good technique for high dimensional problems that accurate result can be obtained with high computational cost. Therefore, it is preferred to pay a little of accuracy to gain a fast algorithm. In such a problem, even if p exceeds the number of samples, there is still enough information for having a good predictive performance. Hence, this technique cannot be a good choice for SNP-data when shrinkage approaches represent poor performance with original dataset.

We also simulated data with $n = 470$ for the dense model with $p = 1000$ in order to evaluate predictive performance with model averaging approach. The posterior probabilities of each model represented in Table 4.2 claim that posterior probability of model is skewed. On the other words, only few models with large m have positive posterior probability and contribute into predic-

tion. Hence, it seems that model averaging only increases computational cost.

Table 4.2: Posterior probability, p.p, of the model with m in $(33, 450)$.

m	356	372	407	410	434	o.w.
p.p	0.04	.01	0.91	.02	0.02	0

Chapter 5

Two-stage Method

Complexity of the genome-wide association studies due to the large p and the small n prevents to achieve a good performance with statistical methods. This motivates a continuing effort to develop two-stage methods (see e.g., Murcray *et al.*, 2009; Zheng *et al.*, 2007). In these kinds of approaches, first a subset of most promising markers is selected for main analysis in the second stage.

In this chapter, we present a new two-stage approach that is a hybrid method of single and simultaneous analyses. In the first-stage, we independently assess the impact of each marker on the complex trait. Then in the second-stage, the markers that met the first-stage threshold are analyzed simultaneously. We develop two models corresponding to two different thresholds. One threshold provides possibility to include marginal and epistatic effects in the model. The other one that is appropriate for the traits with low heritability reduces the risk of missing important effects through first-stage filtering. In these two models, we place a new shrinkage prior, generalized double Pareto (Armagan *et al.*, 2013), on marker effects and obtain all conditional distributions for Gibbs sampling scheme. These new prior specifications for mixed models lead to good predictive performance.

5.1 Method

5.1.1 First-Stage

The strategy for the first-stage is to rank SNPs by measuring the impact of each marker on the complex trait. To assess the association of each marker at a time, we consider a linear regression model with presence of random effects as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\psi}\boldsymbol{\beta} + \mathbf{u} + \epsilon \tag{5.1}$$

$$\boldsymbol{\epsilon} \sim N(0, \tau^{-1}R), \qquad \mathbf{u} \sim N(0, \tau_u^{-1}K)$$

where $\boldsymbol{\psi}$ is a $p \times p$ diagonal matrix with *zero* and *one* entries; if ith predictor is in the model, the ith diagonal entry is *one*. Here, R and K are heterogeneous-residual variance and relatedness matrix respectively.

Although any methods in Chapter 2 can be applied to list markers by their impact on the complex trait, here we screen all markers through Bayesian approach. In this stage, usual conjugate priors can be placed on $\boldsymbol{\beta}$, τ and τ_u since in each time one SNP is in the model. In particular, we consider

$$\boldsymbol{\psi}\boldsymbol{\beta} \mid \tau \sim N(0, \tau^{-1}\Sigma_{\beta_\psi}), \quad \tau \sim \text{Gamma}(a_1, b_1) \quad \tau_u \sim \text{Gamma}(a_2, b_2). \tag{5.2}$$

where $\Sigma_{\beta_\psi} = \text{diag}\{\eta_j\}$. Then, we rank SNPs through

$$ML_1/ML_0$$

which is the odd of presence of each SNP in the model. Here, the ML denotes marginal likelihood of the model; the indexes of the MLs represent the number of predictors in the model. In general, marginal likelihood is defined as

$$\int L(\boldsymbol{\theta} \mid \mathbf{y})\pi(\boldsymbol{\theta})d\boldsymbol{\theta},$$

where $L(\boldsymbol{\theta} \mid \mathbf{y})$ is the likelihood function and $\pi(\boldsymbol{\theta})$ is prior distribution specified on the set of parameters in the model. Hence, we face high dimension

integrals due to presence of random effects in the model. By integrating out parameters $\boldsymbol{\beta}$ and random effects $\mathbf{u}$, we obtain

$$\int_{\tau}\int_{\tau_u}(2\pi)^{-n/2}\frac{b_1^{a_1}b_2^{a_2}}{\Gamma(a_1)\Gamma(a_2)}\tau^{a_1-1}\tau_u^{a_2-1}\left|\frac{A}{\tau}+\frac{K}{\tau_u}\right|^{-1/2}\exp(-\tau b_1-\tau_u b_2)$$
$$\times\exp\left[-\frac{1}{2}y^T\left(\frac{A}{\tau}+\frac{K}{\tau_u}\right)^{-1}y\right]d\tau_u d\tau$$

where $A = K + \mathbf{x}\boldsymbol{\psi}\Sigma_{\beta_\psi}\boldsymbol{\psi}\mathbf{x}^T$. These integrals are intractable, so the closed form of the marginal likelihood is not available. Therefore, we approximate the ML via Laplace method. In order to apply the Laplace approximation, we rewrite the integral as

$$\int_{\tau}\int_{\tau_u}\exp(-nh(\tau,\tau_u))d\tau d\tau_u,$$

where $-nh(\tau,\tau_u)$ is logarithm of the function under the integrals. If $h(\tau,\tau_u)$ is smooth with local minimum $\hat{\tau}$ and $\hat{\tau}_u$ in the interior of $(0,\infty)$, the approximation of the ML is

$$\frac{2\pi}{N}\exp(-Nh(\hat{\tau},\hat{\tau}_u))\mid H(\hat{\tau},\hat{\tau}_u)\mid^{-1/2}+O(1/N).$$

Here, $H(\hat{\tau},\hat{\tau}_u)$ is the Hessian matrix of h.

Because the Laplace approximation is based on a linear Taylor series approximation, it requires certain regularity conditions. However, these conditions fail when the mode lies on the boundary or close to the boundary (Hasio, 1997; Erkanli, 1994). The approximation of the ML can be problematic because of $\tau > 0$ and $\tau_u > 0$ restrictions. To prevent this problem, we parameterize $h(\tau,\tau_u)$ by log transformation of τ and τ_u. This ensures that the parameter space is unrestricted and so the mode is not on the boundary. Hence, we can expect an accurate approximation.

Meanwhile, due to the calculation complexity of $H(\hat{\tau},\hat{\tau}_u)$, we have to use the numerical algorithm. For instance, the first order derivative of the function $h(\tau,\tau_u)$ respect to τ is

$$\frac{\partial h}{\partial\tau}=\frac{b_1}{2}-\left(\frac{a_1}{2}-1\right)\frac{1}{\tau}+\frac{1}{2}\text{trac}\left[\left(\frac{A}{\tau}+\frac{K}{\tau_u}\right)\frac{A}{\tau^2}\right]$$
$$+\frac{1}{2}y^t\left(\frac{A}{\tau}+\frac{K}{\tau_u}\right)^{-1}\frac{A}{\tau^2}\left(\frac{A}{\tau}+\frac{K}{\tau_u}\right)^{-1}y.$$

Evaluating odds of presence of each SNP in the model provides a list of ranking markers in terms of association. Selecting the most promising markers based on the typical threshold 10 is inappropriate in the GWAS; although it is interpreted as a strong evidence of association in many scientific applications (Jeffreys, 1961). This threshold does not serve our purpose for reducing the dimension, while it provides a long list of associated SNPs. Hence, we consider two different thresholds, one is a typical threshold in single marker analysis and the other one is defined based on the safe upper limit of the number of predictors in the second-stage model, which is approximated through simulation study.

5.1.2 Second-stage

In the second stage, we consider different models corresponding to two different thresholds:

- Considering 10^5 as the threshold
  This threshold is the typical threshold in single marker analysis in order to select the SNPs with high posterior odds of presence in the model. The posterior odds is given by

$$\text{PosteriorOdds} = \left(\frac{ML_1}{ML_0}\right) \text{PriorOdds}.$$

  As we have seen in the Chapter 2, in single marker analysis the prior odd of association is very small; however, it can be true only for the problems with few numbers of large-effect markers. Therefore, having high posterior odds requires the ML_1/ML_0 to be large enough in order to overcome the low prior odds.

  By this choice of threshold, the number of selected markers is usually smaller than the number of individuals. This brings the possibility to have the corporation of epistatic or interaction effects for better understanding the nature of genetic and obtaining a more complete picture of complex biological systems.

Note that without applying two-stage approach, genome-wide association studies are $p \gg n$ problems. Hence, searching for all possible pairwise interaction faces practical difficulties due to the large number of pairwise comparisons. For instance, a small set of data in the GWAS contains 100,000 SNPs that approximately entail 4.5×10^9 pairwise interaction. This motivates many studies based on multistage approach for selection and prediction in genetic problem (see, e.g., Evans *et al.*, 2006; Hoh *et al.*, 2000).

After selecting SNPs that met the first-stage threshold, we consider pairwise-interaction of those markers in the model and we call it epistatic model. For these kinds of models, using a single shrinkage prior to control the overall complexity of the model would not be appropriate, because there are many potential for interaction effects to be zero. Hence, we consider a model as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\psi}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma} + \mathbf{u} + \boldsymbol{\epsilon},$$

$$\boldsymbol{\epsilon} \sim N(0, \tau^{-1}R), \qquad \mathbf{u} \sim N(0, \tau_u^{-1}K)$$

where $\boldsymbol{\psi}$ has nonzero diagonal entries equal to the number of selected SNPs through the first-stage. The q vector $\boldsymbol{\gamma}$ contains all possible pairwise-interaction and $\mathbf{Z}$ is their $q \times q$ design matrix.

We place following priors on selected marginal effects, $\boldsymbol{\beta}_\psi$, and their pairwise interaction effects:

$$\boldsymbol{\beta}_\psi \sim N(0, \tau^{-1}\Sigma_{\beta_\psi}), \quad \boldsymbol{\gamma} \sim N(0, \tau^{-1}\Sigma_\gamma)$$

where $\Sigma_{\beta_\psi} = \text{diag}\{\eta_j\}, j = 1, ..., p_s$ and $\Sigma_\gamma = \text{diag}\{\delta_k\}, k = 1, ..., q$.

We give the double-exponential prior to local parameters in the second level of hierarchy such that the hyperparameters are defined locally as

$$\eta_j \sim \exp(\xi_j^2/2), \quad \delta_k \sim \exp(\zeta_k^2/2).$$

Instead of presetting values for ξ_js and ζ_ks, it is appealing to assign prior distributions to these parameters as

$$\xi_j \sim \text{Gamma}(c_1, d_1), \quad \zeta_k \sim \text{Gamma}(c_2, d_2). \tag{5.3}$$

This level of hierarchy automatically accounts for the uncertainty of ξ_js and ζ_ks which affect the rate of shrinkage on each regression coefficients. These 3-level hierarchical priors on marginal and epistatic effects are hierarchical representation of the generalized double Pareto (Armagan *et al.*, 2013). In addition, the prior specification of parameters τ and τ_u are the same as the priors in (5.2) in the first-stage.

The parameters of this model can be estimated by sampling from their full conditional posterior densities through MCMC algorithm. The full conditional posterior densities are given in Section 5.2.

- A threshold that leads to $p_s = 2n$
  In many problems, the complex traits are affected by large numbers of small-effect markers. In this kind of problems, independent screening has low power for truly identifying the most promising SNPs. Therefore, 10^5 may not be a good choice of the threshold. To reduce the risk of missing important SNPs through first-stage screening, we define the threshold equal to the safe upper limit of the numbers of predictors in the second-stage model. This upper limit has been approximated through our simulation studies in previous chapters. While the simulation studies have shown a good predictive performance for at most $2n$ numbers of predictors in the model, we find it as a good choice of the threshold.

  The model in this stage is the same as the model in (5.1) where $\boldsymbol{\psi}$ has $2n$ nonzero diagonal entries equal to the number of selected SNPs through the first-stage. Since in this model $p_s > n$, we give 3-level hierarchical shrinkage prior to selected marker effects, $\boldsymbol{\beta}_\psi$, to avoid over fitting problem. In the first level, prior specifications on parameters of the

model is the same as (5.2). In the second and third levels, we consider

$$\eta_j \sim \exp(\xi_j^2/2), \quad \xi_j \sim \text{Gamma}(c, d). \tag{5.4}$$

The predictive performance of this model is evaluated in the last chapter by estimating parameters of the model through Gibbs sampling scheme. The full conditional posterior densities required for Gibbs algorithm are presented in section 5.2.

5.2 Full Conditional Posterior Densities

To simplify notation for the full conditional posterior densities, let X_ψ denotes an $n \times p_s$ matrix such that each vector is corresponded to a selected marker in the first-stage.

- Full conditional posterior densities for the epistatic model
  $\boldsymbol{\beta}_\psi \mid \mathbf{y}, \mathbf{u}, \boldsymbol{\gamma}, \tau, \Sigma_{\beta_\psi} \sim N\left(\boldsymbol{\mu}^p_{\beta_\psi}, \Sigma^p_{\beta_\psi}\right),$
  $\boldsymbol{\gamma} \mid \mathbf{y}, \boldsymbol{\beta}_\psi, \mathbf{u}, \tau, \Sigma_\gamma \sim N\left(\boldsymbol{\mu}^p_\gamma, \Sigma^p_\gamma\right),$
  $\mathbf{u} \mid \mathbf{y}, \boldsymbol{\beta}_\psi, \boldsymbol{\gamma}, \tau, \tau_u \sim N\left(\boldsymbol{\mu}^p_u, \Sigma^p_u\right),$
  $\tau \mid \mathbf{y}, \boldsymbol{\beta}_\psi, \boldsymbol{\gamma}, \mathbf{u}, , \Sigma_{\beta_\psi}, \Sigma_\gamma \sim \text{Gamma}\left(a_1^p, b_1^p\right),$
  $\tau_u \mid \mathbf{u} \sim \text{Gamma}\left(a_2 + n/2, \quad \mathbf{u}^T K^{-1}\mathbf{u} + b_2\right),$
  $\xi_j \mid \gamma_j, \tau \sim \text{Gamma}\left(c_1 + 1, \quad \tau^{1/2} \mid \gamma_j \mid + d_1\right),$
  $\eta_j^{-1} \mid \gamma_j, \xi_j, \tau \sim \text{IN-Gaussian}\left(\left(\frac{\xi_j^2}{\tau\gamma_j^2}\right)^{1/2}, \xi_j^2\right),$
  $\zeta_k \mid \gamma_k, \tau \sim \text{Gamma}\left(c_2 + 1, \quad \tau^{1/2} \mid \gamma_k \mid + d_2\right),$
  $\delta_k^{-1} \mid \gamma_k, \zeta_k, \tau \sim \text{IN-Gaussian}\left(\left(\frac{\zeta_k^2}{\tau\gamma_k^2}\right)^{1/2}, \zeta_k^2\right)$
  where
  $\boldsymbol{\mu}^p_{\beta_\psi} = \left(\mathbf{X}_\psi^T R^{-1}\mathbf{X}_\psi + \Sigma^{-1}_{\beta_\psi}\right)^{-1}\left(\mathbf{X}_\psi^T R^{-1}(\mathbf{y} - \mathbf{Z}\boldsymbol{\gamma} - \mathbf{u}) + \Sigma^{-1}_{\beta_\psi}\mu\right),$
  $\boldsymbol{\mu}^p_\gamma = \left(\mathbf{Z}^T R^{-1}\mathbf{Z} + \Sigma^{-1}_\gamma\right)^{-1} Z^T R^{-1}\left(\mathbf{y} - \mathbf{X}_\psi\boldsymbol{\beta}_\psi - \mathbf{u}\right),$
  $\boldsymbol{\mu}^p_u = \left(K^{-1}\tau_u + \tau R^{-1}\right)^{-1} R^{-1}\left(\mathbf{y} - \mathbf{X}_\psi\boldsymbol{\beta}_\psi - \mathbf{Z}\boldsymbol{\gamma}\right)$
  $\Sigma^p_{\beta_\psi} = \tau^{-1}\left(\mathbf{X}_\psi^T R^{-1}\mathbf{X}_\psi + \Sigma^{-1}_{\beta_\psi}\right)^{-1},$
  $\Sigma^p_\gamma = \tau^{-1}\left(\mathbf{Z}^T R^{-1}\mathbf{Z} + \Sigma^{-1}_\gamma\right)^{-1},$

$\Sigma_u^p = (K^{-1}\tau_u + \tau R^{-1})^{-1}$,
$a_1^p = \dfrac{n + p_s + q}{2} + a_1$,
$b_1^p = \dfrac{1}{2}\Bigg((\mathbf{y} - \mathbf{X}_\psi \boldsymbol{\beta}_\psi - \mathbf{Z}\boldsymbol{\gamma} - \mathbf{u})^T R^{-1} (\mathbf{y} - \mathbf{X}_\psi \boldsymbol{\beta}_\psi - \mathbf{Z}\boldsymbol{\gamma} - \mathbf{u}) + \gamma^T \Sigma_\gamma^{-1} \gamma + \boldsymbol{\beta}_\psi^T \Sigma_{\beta_\psi}^{-1} \boldsymbol{\beta}_\psi \Bigg) + b_1$.

Here, we obtained density of $\xi_j \mid \gamma_j, \tau$ as conditional posterior of ξ_j instead of $\xi_j \mid \eta_j$. As we have

$$\pi(\xi_j \mid \gamma_j, \tau) \propto \pi(\gamma_j \mid \xi_j, \tau)\pi(\tau)\pi(\xi_j),$$

to obtain$\pi(\gamma_j \mid \xi_j, \tau)$, we integrate out η_j

$$\begin{aligned} \pi(\gamma_j \mid \xi_j, \tau) &= \int \pi(\gamma_j \mid \eta_j, \tau)\pi(\eta_j \mid \xi_j) d\eta_j \\ &= \exp\Bigg(- \xi_j \tau^{1/2} \mid \gamma_j \mid \Bigg) \frac{\xi_j}{2} \tau^{1/2}. \end{aligned}$$

- Full conditional posterior densities for $2n$ top-ranking markers
$\boldsymbol{\beta}_\psi \mid \mathbf{y}, \mathbf{u}, \tau \sim N\left(\boldsymbol{\mu}_{\beta_\psi}^p, \Sigma_{\beta_\psi}^p\right)$
$\mathbf{u} \mid \mathbf{y}, \boldsymbol{\beta}_\psi, \tau, \tau_u \sim N\left(\boldsymbol{\mu}_u^p, \Sigma_u^p\right)$
$\tau \mid \mathbf{y}, \boldsymbol{\beta}_\psi, \mathbf{u}, \boldsymbol{\mu}, \Sigma_{\beta_\psi} \sim \text{Gamma}\left(a_1^p, b_1^p\right)$,
$\tau_u \mid \mathbf{u} \sim \text{Gamma}\left(a_2 + n/2, \quad \mathbf{u}^T K^{-1} \mathbf{u} + b_2\right)$,
$\xi_j \mid \beta_{\psi_j}, \tau \sim \text{Gamma}\left(c + 1, \quad \tau^{1/2} \mid \beta_{\psi_j} \mid + d\right)$.
$\eta_j^{-1} \mid \beta_{\psi_j}, \xi_j, \tau \sim \text{IN-Gaussian}\left(\left(\dfrac{\xi_j^2}{\tau \beta_{\psi_j}^2}\right)^{1/2}, \xi_j^2\right)$.
where
$\boldsymbol{\mu}_{\beta_\psi}^p = \left(\mathbf{X}_\psi^T R^{-1} \mathbf{X}_\psi + \Sigma_{\beta_\psi}^{-1}\right)^{-1} X_\psi^T R^{-1} (\mathbf{y} - \mathbf{u})$,
$\boldsymbol{\mu}_u^p = \left(K^{-1}\tau_u + \tau R^{-1}\right)^{-1} R^{-1} \left(\mathbf{y} - \mathbf{X}_\psi \boldsymbol{\beta}_\psi\right)$
$\Sigma_{\beta_\psi}^p = \tau^{-1} \left(\mathbf{X}_\psi^T R^{-1} \mathbf{X}_\psi + \Sigma_{\beta_\psi}^{-1}\right)^{-1}$,
$\Sigma_u^p = (K^{-1}\tau_u + \tau R^{-1})^{-1}$,
$a_1^p = \dfrac{n + p_s}{2} + a_1$,
$b_1^p = \dfrac{1}{2}\Bigg((\mathbf{y} - \mathbf{X}_\psi \boldsymbol{\beta}_\psi - \mathbf{u})^T R^{-1} (\mathbf{y} \mathbf{X}_\psi \boldsymbol{\beta}_\psi - \mathbf{u}) + \beta_\psi^T \Sigma_{\beta_\psi}^{-1} \beta_\psi \Bigg) + b_1$.

5.3 Discussion

While simultaneous analysis based on shrinkage priors have a limitation of disparity between the number of predictors and the number of samples, reducing the dimensionality of the data is required. We present a new two-stage approach that is a hybrid method of single marker analysis and simultaneous analysis. In the first-stage, we select the most promising SNPs by assessing the association of each SNP independently. We measure the association through the odd of presents of each SNP in the model, which is common in Bayesian single-based analysis. To select SNPs from the list of ranking SNPs in the first-stage, we consider two different thresholds, one appropriate for very sparse problems and the other for problems with large numbers of small-effects. Respectively, we develop two different models for problems with related samples. In these two models, we place generalized double Pareto as shrinkage prior on marker effects. The parameter of the models are estimated through Gibbs sampling schemes.

Chapter 6

Application

Genomic-enabled prediction is becoming increasingly important in animal and plant breeding and it also receiving attention in human genetics. Prediction of genetic values early in life leads to select the animals or plant with high quality of desired products or traits and in human leads to diagnose disease susceptibility. Achieving accurate prediction has been introduced a big challenge to the statistical analysis as we have discussed in previous chapters.

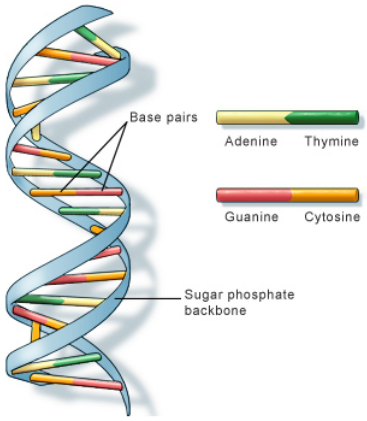
In this chapter, we represent our real SNP-data analysis. In the first section, we introduce the problem. Section 2 to 4 are devoted to some preliminary analyses, quality control procedure for SNP-data, dimension reduction by clustering SNPs via linkage disequilibrium and detecting population structure. In section five, we attempt to identify the associated SNPs based on single marker analysis. Finally in section 6, we applied the proposed methods in Chapter 5 and evaluated the prediction performance via a comparison with two other prior specifications.

6.1 Dataset

The real SNP-data set comes from an animal breading research project. The aim of the research is to improve the quality and quantity of the milk production of cattle. The data contains $707,962$ SNPs genotyped for 607 numbers of Holestein Bulls.

6.1.1 SNPs as Predictors

The aim of genetic studies is to capture the information in DNA related to complex traits or diseases. DNA is the genetic material determining the makeup of all living cells and many viruses. It consists of two long strands of nucleotides linked together in a structure resembling a ladder twisted into a spiral. The information in DNA is stored as a code made up of four chemical bases: adenine (`A`), guanine (`G`), cytosine (`C`), and thymine (`T`). Adenine always pairs with thymine, and cytosine pairs with guanine.



Four Chemical Bases

Adenine (A)
Guanine (G)
Cytosine (C)
Thymine (T)

The order of these bases in genomic sequence determines the information available for building and maintaining an organism, similar to the way in which letters of the alphabet appear in a certain order to form words and sentences. The most part of DNA sequence is similar among members of a biological species. For instance, human DNA consists of about 3 billions bases such that more than 99 percent of those bases are the same in all people. Therefore, in genetic problems we attempt to understand the impact of genetic variants in complex traits. To this end, those single nucleotides or bases (`A, T, C` or `G`) in the genome which are different between members of a biological species so called single nucleotide polymorphisms, SNPs, are assayed. The Figure 6.1 represents 3 SNPs in 5 DNA sequences.

Since each marker can be assayed as {`T, A`} or {`C,G`} in two-locus studies,

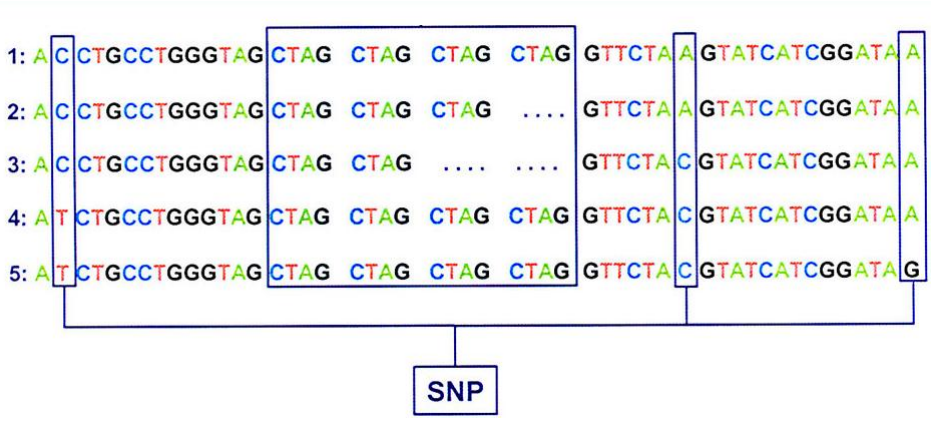


Figure 6.1: Close-up view of DNA sequences and SNPs in.

three combinations of genotypes can be considered for each SNP; in a simplest way without considering the name of the bases, these are `BB`, `Bb`, or `bb`. To convert the genotyped SNPs to count variables, let assume `b` is minor allele frequency, which is referred to the least frequent allele in a given population. Then, for additive model that we have focused on, each SNP represent the numbers of `b` for each sample as

$$\begin{cases} 0 & \text{if the genotype of the SNP is } \texttt{BB} \\ 1 & \text{if the genotype of the SNP is } \texttt{Bb} \\ 2 & \text{if the genotype of the SNP is } \texttt{bb}. \end{cases}$$

6.1.2 Phenotype

In the case of genomic prediction for complex traits of animals, the response might be single or repeated measure of individuals' phenotypic performance, information on progeny, estimated breeding values (EBV) from genetic evaluations or a pooled mixture of more than one of these information sources. Our study is based on EBV that is an estimate of breeding value. Breading value (BV) is the genetic value of an individual determined by the mean value of its progeny; i.e., it is the genetic transmitting ability from a generation to the next. This is an efficient way to combine heritability information with performance of relatives and progeny to predict breeding value.

6.2 Quality Control

Data cleaning is the first and essential step for data analysis. Whether the goal is prediction of the outcomes or to discover new biology underlying the trait of interest, the inference of GWAS depends upon the overall quality of the data. Even simple statistical tests of association are compromised in the context of GWAS with data that have not been properly cleaned, potentially leading to false-negatives and false-positive associations. Hence, we followed the common steps for quality control in genome-wide association studies to prevent these problems.

6.2.1 Cleaning Data over SNPs

SNPs Signed to Chromosome *zero*

Before checking genotype quality of SNPs, it needs to assure that SNPs are assigned to specific chromosome. It is usual to have some SNPs not aligned to the current genome assembly. Therefore, they signed to the chromosome *zero*. These SNPs must be removed from the data set. In our dataset, $2,078$ numbers of SNPs have been recorded for chromosome *zero*.

Call Rate

The proportion of a genotype call for each marker, genotyping efficiency, is a good indicator of marker quality. The SNPs' assays that failed on a large number of samples are poor assays, and are likely to result in spurious data. Hence, SNPs with low call rate must be discarded. A recommended threshold for removing SNPs with low call rate is approximately $90 - 99\%$, although this threshold may vary from study to study and it should be decided by researcher. We excluded $19,084$ SNPs from the dataset based on %90 threshold. But after doing some analysis, we found %99 a better choice for our problem, which discarded $38,900$ numbers of SNPs from the data. Table 6.1 presents the call rate summary of the genotyped SNPs.

Table 6.1: Summary table of genotype call rate across samples.

	$X <= .9$	$.9 < X <= .98$	$.98 < X <= .99$	$X > 99$
No	19084	62619	38900	$775,884$
Prop	0.025	0.08	0.05	0.845

Minor Allele Frequency

Another important issue in quality control is to exclude SNPs with low variability for minor allele so called rare SNPs. This filtering step helps to improve statistical power. So, removing extremely rare SNPs including any monomorphic SNPs has been recommended. The choice of threshold depends on the size of study and the impact of SNP-effects in *priori*. In our study, we removed $191,936$ numbers of SNPs with %5 threshold.

Hardy-Weinberg Equilibrium

Checking for Hardy-Weinberg Equilibrium (HWE) is the final step in the quality control analysis of markers in genome-wide association studies. Under Hardy-Weinberg assumptions, allele and genotype frequencies can be estimated from one generation to the next. Typically, HWE deviations toward an excess of heterozygotes reflect a technical problem in the assay, such as non-specific amplification of the target region. If no technical errors are detected then a number of biologically plausible explanations exist for HWE deviations such as population stratification, assortative mating and inbreeding. In animal studies and some human population, Hardy-Weinberg equilibrium check may not be as usual due to inbreeding and nonrandom mating in the sample population. Non random mating and inbreeding are two conditions that violate crucial assumption of HWE because inbreeding increases the frequency of homozygous, and decreases the frequency of heterozygous genotypes. In our data, samples in the same farms are likely to share the same alleles, inherited from common ancestors. Therefore, their

progeny has an increased chance of being autozygous that refers to inherit a copy of exactly the same ancestral allele from both parents. In our analysis, 81,720 numbers of SNPs have shown departure from HWE with 0.1 threshold.

6.2.2 Missing Value

Imputation based on Pedigree and Probability

In family based study, using pedigree for imputation is more reliable, especially in our case that inbreeding increases the rate of homozygous SNP. High rate of homozygosity reduces the number of possible combination of genotypes that can be inherit by children and consequently increases the probability of occurrence of each combination. However this puts pedigree at the top of the imputation methods, it can be applied only when the genome of parents have been genotyped. While in our data set a few numbers of paternal and maternal genotypes are available, we cannot impute missing SNPs based on pedigree.

Imputation based on Linkage Disequilibrium

Markers in the same chromosome are in linkage disequilibrium, LD, that vanishes by genetic or physical distance between SNPs. It is therefore desirable to develop a flexible imputation approach that takes into account the LD in neighboring SNPs. A variety of techniques has been applied to the problem of imputing missing genotypes. A common statistical approach is to infer missing genotypes from haplotype frequencies of population samples. More recent approaches incorporate models of recombination by partitioning markers into haplotype blocks based on entropy measures or by inferring a mosaic of haplotype clusters. Tree-based imputation methods have been also developed that impute missing genotypes on the basis of perfect phylogeny rather than haplotype structure.

6.2.3 Cleaning over Samples

In the GWAS, the sample size is usually very small respect to the numbers of SNPs. Therefore, cleaning over samples and detecting for the structure behind the data must be done very carefully because each sample plays significant role in the analysis.

Call Rate

A large proportion of SNP assays failing on an individual DNA sample may indicate a poor quality DNA sample, which could lead to aberrant genotype calling. Samples with low genotyping efficiency, or call rate, should be eliminated from analysis. The recommended threshold is $98 - 99\%$ for excluding low call rate SNPs over samples after removing low genotyped SNPs across samples. This threshold is an approximate threshold so the exact threshold may vary from study to study depending on the used platform for genotyping, quality of the DNA samples, the variability in population and equipment error. The threshold should be determined based on a goal whereby a balance between minimizing the number of samples dropped and maximizing genotyping efficiency is attained.

Table 6.2: Summary table of genotyped call rate over samples.

	$X <= .9$	$.9 < X <= .98$	$.98 < X <= .99$	$X > .99$
No	40	57	222	328
Prop	0.07	0.09	0.37	0.54

By looking at the Table 6.2, we can realize that the recommended threshold is not appropriate for our data set. If we eliminate the samples with 99%, we will lost almost 50% of the samples and with 98%, 57 samples. So we decided to keep most of the samples in the dataset by threshold 90%. Hence, 40 samples have been excluded.

Quality control of SNP-data is very intensive from computational point of

view since dataset is very large. Thus, it cannot be done with any software like `R`. To do this we used `Plink` software that is a fast software built for SNP-data. After quality control, our dataset includes 555,651 SNPs with 567 observations.

6.3 Reducing Dimension via Linkage Disequilibrium

Many studies have been shown that SNPs in the genome have groups of neighbors such that they are all nearly perfectly correlated with each other due to the LD. The other words, the genotype of one SNP can perfectly predict those of correlated neighboring SNPs. These segments of SNPs in high linkage disequilibrium in animals are longer than human. One SNP can thereby serve as proxy for many others in analysis. By considering this fact, we can reduce dimensionality of the problem.

We applied hierarchical clustering approach in order to detect the correlated neighboring SNPs. Although this clustering did not seem to be problematic, we could not simply apply the standard algorithm because of the large numbers of objects, SNPs. To overcome this problem related to calculating similarity matrix, we defined a window with length of 200 base pairs that moved with step size of 20 base pairs. The length of window was based on physical distance while Centimorgan distance could not be calculated before data analysis. After obtaining the similarity matrix, we clustered the SNPs with at least 85% correlation in each cluster. Then among all SNPs in each cluster, the one that was closest to the others were tagged. This procedure selected 135,545 numbers of SNPs for the main analysis.

6.4 Population Structure

After quality control, a major practical issue for studying complex traits or disease is to identify population structure in the data while ignoring this step reduces the power of genetic studies. Structure in the data might be

caused by cryptic relatedness or population stratification. Cryptic relatedness refers to presence of unknown genetic relationships between individuals within the study samples. Population stratification occurs when the study samples comprise multiple groups of individuals who differ systematically in both genetic ancestry and the phenotype under investigation. If we do not account for population structure, we will identify spurious associations due to differences in ancestry rather than true association of alleles to the traits. Thus, it is critical to check for population structure within the samples in order to avoid false discoveries and bias in prediction.

6.4.1 Principal Components Analysis

Principal component-based methods applied to genotypes provide information about population structure, and have been widely used to correct for the stratification. Typically in the PC analysis, if the first few PCs capture most of the variation in the data, we have population stratification. In order to account for the stratification, we need to add those PCs to the model as extra covariates.

However the use of PCA in genetics can be dated to several decades ago before the advent of SNP-data, it may be faced more challenges in the GWAS. In these kinds of problems, predictors or SNPs are in linkage disequilibrium in genetics regions. Hence, by applying the PCA directly on the whole dataset, the first principal component may simply reflect unusually stretches of LD rather than population stratification. To avoid this problem, we first thinned the data by LD in order to make a set of SNPs that are almost uncorrelated. It should be emphasized that the aim of PCA is to identify population structure not dimension reduction. Therefore, it differs with the goal of applying the sparse principal methods (see, e.g., Zou *et al.*, 2006) for reducing the numbers of variables in each components.

To apply PCA, we first clustered SNPs by LD with similar algorithm in Section 6.3. In order to select a subset of approximately independent SNPs, the correlation between each cluster is considered to be less than

4% (Lee *et al.*, 2012). This provided 37,916 numbers of clusters. Then, the closest SNPs to all other SNPs within each cluster selected for PCA analysis. After applying the PCA, we faced with tiny components such that the first component explained around 1% of the variation in the data. Thus, based on this analysis, the dataset does not stratify to different populations.

6.4.2 Identical by Descent

Study of relationship between samples is phrased in terms of probabilities that a set of genes have descended from a single ancestral gene. This criterion is the probability that individuals are identically-by-decent, IBD. Hence, two individuals are said to be related if the allele or alleles of one are IBD to those of the other(Weir, et al., 2006).

Relatedness analysis through IBD depends on the pedigree structure. This common ancestor may be a parent, grandparent, etc. Since in our dataset most of the parental and maternal genetic information is missed and imputation introduces huge bias in this analysis, we found it inadequate to apply.

6.5 Single Marker Analysis

In this stage, we aimed to identify the truly associated SNPs with longevity trait through single-based approaches, which we have seen in Chapter 2.

6.5.1 Linear Regression Model

In single marker analysis, we first consider a linear regression to assess the impact of each SNP on the trait at the time. After fitting the model for each SNP, we calculated all the p-values based on Wald test for multiple comparisons. As it can be seen from the histogram of p-values in Figure 6.2, the left side of the histogram is different with $U(0, 1)$. Furthermore, the histogram density beyond 0.9 looks fairly flat. These can be evidence of

difference between empirical density of p-values and theoretical null density. The same conclusion can be also reached from left panel in Figure 6.4.

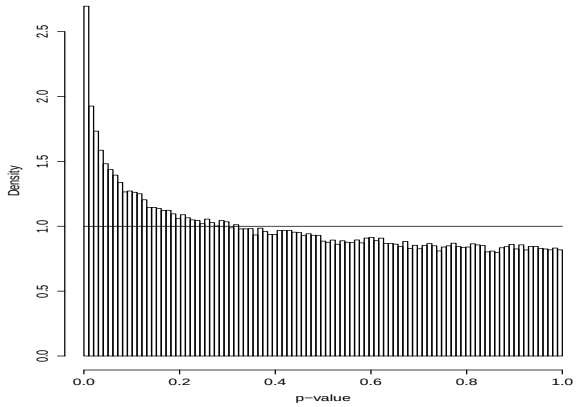


Figure 6.2: p-values histogram of longevity, horizontal line indicates $U(0,1)$.

To select associated SNPs, we first applied false discovery rate criterion via BH algorithm. Unfortunately, based on the BH algorithm, a long list of SNPs were significant that cannot be practical to work with. Then, we simply defined a cutoff threshold based on Bonferoni correction since it is widely used in practice in genome-wide association studies. For level 1% for Type I error, 6 SNPs located in three different chromosomes selected as associated SNPs. Figure 6.3 so called Manhatan plot represents the location of selected SNPs.

Although these selected SNPs seems to be spurious association by looking at left panel in Figure 6.4, we consulted the biologist in research group. Then we found these SNPs are not neither in genes nor close to genes. They are not even in the chromosomes that biologist expected in *priori* for this trait. Therefore, the result seems to be spurious associations.

We repeated our analysis based on z-values. The left panel in Figure 6.5 represents z-values' histogram of our statistical tests. The theoretical null density, standard normal, is drawn with blue dash-line and the empirical density is drown by green solid-line. The difference between these two densities can be seen easily in the figure. The empirical density is very wider

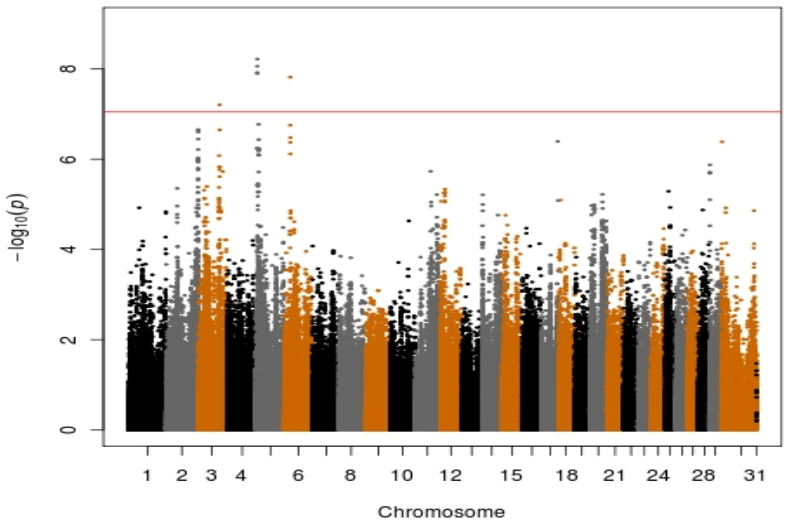


Figure 6.3: Manhatan Plot

than normal standard. Selecting SNPs based on (2.3) provides a long list of SNPs associated with the trait. These SNPs are in the tail areas colored by pink in the figure. To bypass this problem, we estimated null density by

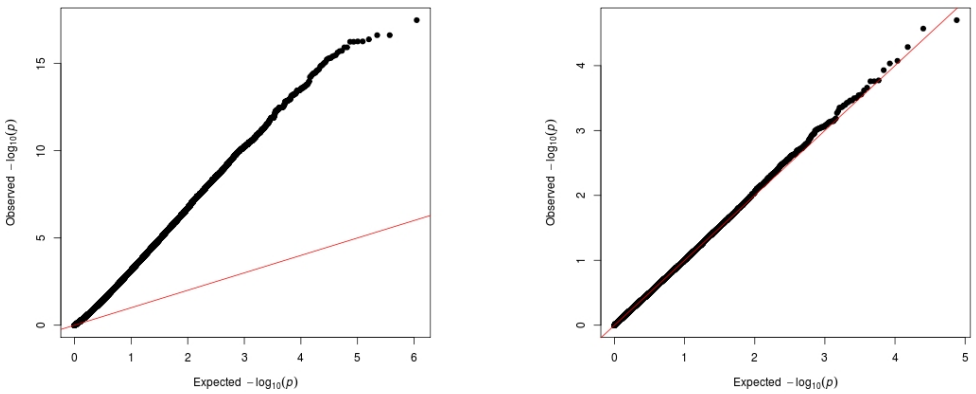


Figure 6.4: Left panel: q-q plot determined from linear regression. Right panel: q-q plot determined from linear mixed model.

central matching and maximum likelihood approaches. Following the central

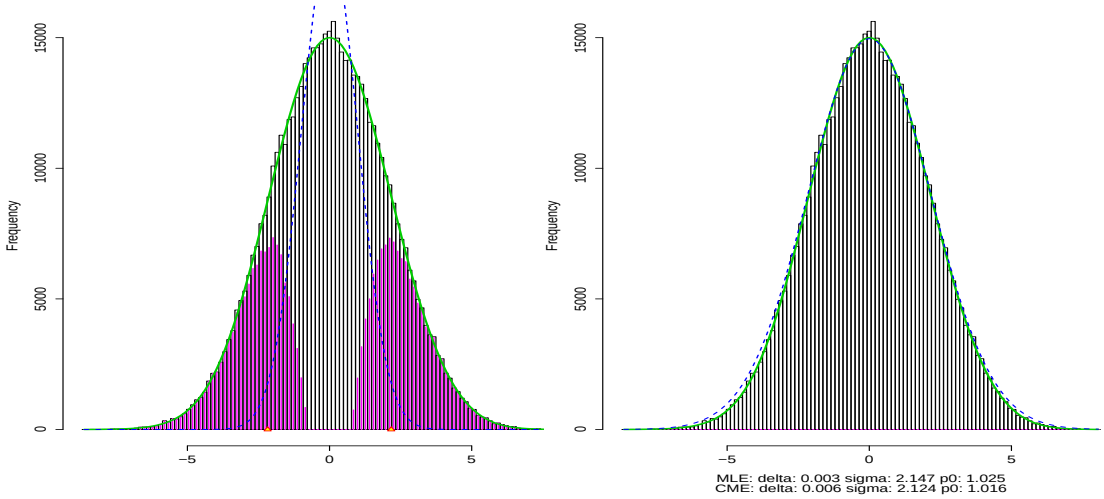


Figure 6.5: Left panel: green solid-line is spline based estimator of $f(z)$, blue dashed-line is $\pi_0 f_0(z)$based on theoretical null distribution $N(0,1)$. Right panel: green solid-line is spline based estimator of $f(z)$, dashed-line is empirical null density.

matching procedure, we estimated

$$(\hat{\delta}_0, \hat{\sigma}_0) = (0.006, 2.124), \quad \hat{\pi}_0 = 1.$$

The estimations based on maximum likelihood approach are

$$(\hat{\delta}_0, \hat{\sigma}_0) = (0.003, 2.147), \quad \hat{\pi}_0 = 1.$$

that is very similar to central matching estimates. The right panel in Figure 6.5 shows the estimated null density and empirical density that are almost the same. In other words, we can say the figure dose not represent mixture of two densities, density of null and non null cases. Thus, we could not detect any SNPs associated with the trait.

6.5.2 Linear Mixed Model

While estimating null density did not help for detecting any SNP associated with trait, we guessed the huge deviation between empirical density and theoretical null density might be the cause of this problem. This deviation typically arises due to population stratification or relatedness among samples.

Since we did not find any stratification in the data through PCA, relatedness samples might be the reason for this deviation. For accounting for related sample, mixed model is a powerful tool. Therefore, this time we considered a linear regression with presence of random effects as

$$\mathbf{y} = \mathbf{x}_i\beta_i + \mathbf{u} + \boldsymbol{\epsilon}, \quad \mathbf{u} \sim N(\mathbf{0}, \tau_u^{-1}K), \quad \boldsymbol{\epsilon} \sim N_n(\mathbf{0}, \tau^{-1}R), i = 1, ..., p,$$

similar to the model (5.1) with one predictor in the model. Relatedness matrix K is calculated from the pedigree such that its entries represents the degree of relatedness between pairs of sample and take values in $[0, 1]$.

We did redo our single-based analysis for mixed model. The right panel in Figure 6.4 is *q-q* plot based on analysis with linear mixed model. As it represents, random effects correct the deviation perfectly; but we still cannot expect to detect any association.

However we did not detect any SNPs associated with the trait, the result of the analysis reveals that the trait is affected by large numbers of small-effect markers and single-based analysis is not powerful to detect weak association. This motivated us to analyze the data through two-stage methods proposed in Chapter 5. Furthermore, we realized that the best model for our data is linear mixed model that accounts for related samples.

6.6 Two-stage Analysis

In this section, we evaluated the performance of our proposed model by ten-fold cross validation in comparison with two other prior specifications, Bayesian ridge, BR, and Bayesian lasso, BL.

6.6.1 Two-stage methods with threshold $2n$

Following our approach that is a two-stage method, we first evaluated odds of presence of each SNP in the model. After ranking the SNPs by their importance based on calculated odds, we selected $2n = 1134$ top-ranking SNPs

for the second-stage analysis.

To select the values for hyperparameters of the second-stage model in (5.2) and (5.4), we first ran a 5-fold cross validation. Ideally, we should test a large set of values for 6 hyperparameters, but this may be time consuming. Therefore, for hyperparameters of marker effects, we considered $d = \sqrt{c+1}$ since this choice ensures having continuity property and creates a trade-off between sparsity and tail-robustnes (Armagan *et al.*, 2013). We ran cross validation for different values of c: 1, 2.5, 3, 3.5. Generally, the rate of shrinkage increases along this path. For the other hyperparameters of the model, we set $(a_1, b_1) = (a_2, b_2)$ as $(0.001, 0.001)$, (0.01, 0.01), (0.1, 0.1), (0.3, 0.3). Table 6.3 presents average of 5 mean square prediction errors, MSPEs, for different values of hyperparameters and the standard deviation from 50 bootstrap samples in subscript.

Table 6.3: Average of 5 out-of-samples MSPEs for different values of hyperparameters and their standard deviations from 50 bootstrap samples in subscript.

$(a_1, b_1) = (a_2, b_2)$	c			
	1	2.5	3	3.5
(0.001,0.001)	$0.698_{(0.018)}$	$0.539_{(0.012)}$	$0.526_{(0.013)}$	$0.529_{(0.010)}$
(0.01, 0.01)	$0.686_{(0.019)}$	$0.528_{(0.011)}$	$0.519_{(0.011)}$	$0.523_{(0.015)}$
(0.1, 0.1)	$0.736_{(.028)}$	$0.564_{(0.018)}$	$0.551_{(.022)}$	$0.557_{(0.023)}$
(0.3, 0.3)	$0.845_{(0.028)}$	$0.576_{(0.018)}$	$0.572_{(0.058)}$	$0.569_{(0.023)}$

It is seen from Table 6.3 that the MSPE for $c = 3$ and $(a_1,\ b_1) = (a_2,\ b_2) = (0.01,\ 0.01)$ is smaller than the other values. Therefore, the cross validation gave $c = 3$ as the best choice for shrinkage parameter and 0.01 for the other parameters.

To compare prediction performance of our model with the BL and the BR, we ran 10-fold cross validation. For the BL, the posteriors are not sensitive to the prior specification (3.5) as long as r and s are small so that the

priors are relatively flat (Park & Casella, 2008; Yi & Xu, 2008). Thus, we set these hyper parameters as small values, 0.1. For setting the hyperparameters of the SNP effects in the BR, we first fixed the value of the shape parameters ν in (3.11) as suggested by Frühwirth-Schnatter & Wagner (2010). Then, we considered different values for scale parameters as $s = 0.001,\ 0.01,\ 0.1$ and found 0.01 as the best choice for s.

Table 6.4 represents average of MSPEs of 10-fold cross validation. The index numbers are the average standard errors of MSPEs obtained by 100 bootstrap samples of 10 MSPEs corresponding to 10 folds. For making better comparison, we obtained the deviance information criterion, DIC. We also calculated average of correlation between predicted and observed values in validation sets. The comparison through MSPEs and DICs shows that our model out performs the other competitors.

Table 6.4: First column: average of 10-fold cross validation MSPEs of the new two-stage methods denoted by GDP, the BR and the BL and their standard deviations based on 100 bootstrap samples in subscript. Second column: average of correlation of observed values and predicted values in 10 validation sets. Third column: the DIC.

Model	MSPE	Cor	DIC
GDP	$0.518_{(0.0276)}$	0.701	581.42
BR	0.551(0.0362)	0.662	729.34
BL	$0.568_{(0.0344)}$	0.653	811.53

The total phenotypic variance for the trait given $\boldsymbol{\beta}$ can be written as

$$\mathrm{V}_y = \sum_{j=1}^{p_s}\sum_{j'=1}^{p_s} \beta_j\beta_{j'}\ \mathrm{cov}(x_j, x_{j'}) + s\tau^{-1} + s_b\tau_u^{-1} \tag{6.1}$$

where s and s_u are the mean of diagonal elements in R and K respectively. In other words, $s = \frac{1}{n}\sum_{i=1}^{n} r_{ii}$ and $s_u = \frac{1}{n}\sum_{i=1}^{n} k_{ii}$ where r_{ij} and k_{ij} are the

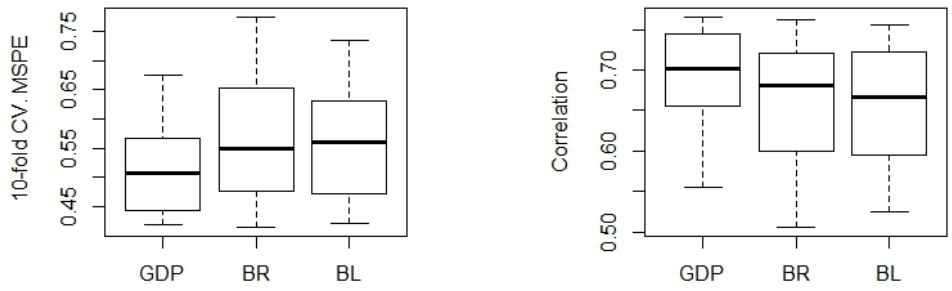


Figure 6.6: Left-side: box plot of MSPE obtained through 10-fold cross validation, right-side: box plot of correlation between predicted and observed values in validation sets.

ijth elements of matrixces R and K. Here, $\text{cov}(x_j, x_{j'})$ is covariance between X_j and $X_{j'}$ if $j \neq j'$ otherwise it is the variance of X_j.

The calculated total phenotypic variance from (6.1) is 1.700. The total genetic variance contributed by the additive effects of the markers calculated from the first term of the right hand side of (6.1) is 0.483. The proportion of the phenotypic variance explained by total genetic variance is called heritability denoted with h^2. As it is clear from (6.1), h^2, accounts for the covariance between markers as well. If we ignore the contribution from the covariance, the proportion of the phenotypic variance explained by each marker can be approximated by

$$h_j^2 = \frac{\beta_j \text{var}(x_j)}{\text{V}_y}. \tag{6.2}$$

To select significant SNPs based on heritability, Hoti & Sillanpaa (2006) suggested to present a threshold value, c, such that one SNP is included in the final model if the heritability explained by this SNP is greater than c. Therefore this threshold can be chosen on more subjective grounds. It is more appropriate in genetic problems, while heritability is different for different complex traits. For instant, the heritability for the milk protein yield is expected to be small due to previous studies. With this knowledge, we make a small change in Hoti & Sillanpaa's strategy. Instead of setting a threshold

on heritability of each SNP, we consider a threshold on total heritability of a set of top SNPs ranked based on heritability. After calculating heritability by substituting the mean of posterior samples of β_js in (6.1), a set of top ranking SNPs with total heritability above 0.2 was selected. The estimated effect sizes and marginal heritabilities of 32 selected SNPs as well as their chromosomes' numbers are tabulated in Table 6.5.

Table 6.5: Position of selected SNPs with their effect sizes and heritebilities

index	Chr	β	$h^2(\%)$	index	Chr	β	$h^2(\%)$
27	1	-0.0562	0.131	620	13	-0.0578	0.150
40	1	0.0617	0.192	624	13	0.0585	0.173
54	1	-0.0983	2.930	631	13	0.0569	0.138
82	2	-0.0580	0.120	652	14	-0.0595	0.232
116	3	-0.0572	0.156	770	17	0.0905	2.552
122	3	-0.0572	0.138	803	18	0.0569	0.140
151	4	-0.0595	0.204	813	18	0.0790	1.584
180	4	-0.0591	0.206	950	22	0.0586	0.179
192	4	0.0796	1.529	1032	25	0.0715	0.976
293	6	0.0707	0.906	1039	25	-0.0753	1.091
332	7	-0.0726	0.950	1045	25	-0.0553	0.110
371	8	0.0571	0.136	1103	28	-0.0503	0.100
387	8	0.0606	0.798	1106	28	0.0617	0.741
420	9	0.0636	0.821	1111	28	0.0691	0.880
430	10	0.0566	0.140	1114	28	0.0589	0.197
571	12	-0.0581	0.182	1133	29	0.0580	0.200

Among these selected SNPs 19 markers out of 32 makers have been found in the known genes or close to them. Figure 6.7 shows pieces of chromosome 1 and 10 as examples. The red regions in these pictures indicate the locations of the genes associated with milk protein yield that have been found through previously studies. The two vertical red lines represent the locations of two selected markers in our study. As it is shown in the figure, the selected SNP in Chromosome 1 is in gene `PPP2R3A` and the one in Chromosome 10 is

between genes SAV1 and NIN. The other markers, that are not located close to known regions in the genome, can be sight to regions that have potential to be identified as novel-genes if the future studies also detect significant markers near those locations.

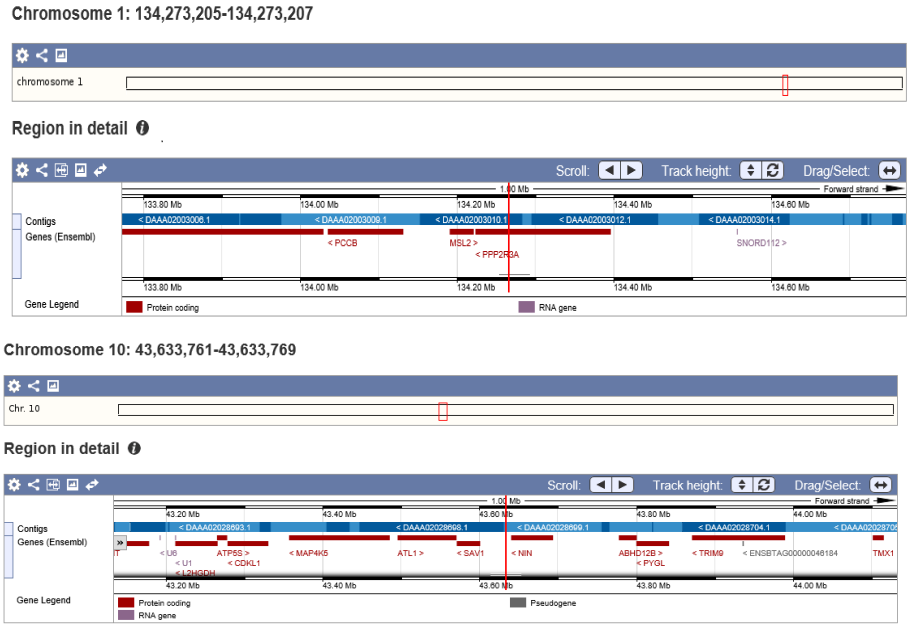


Figure 6.7: Location of two selected markers from chromosome 1 and 10.

We further examined the correlations between the SNPs that are from the same chromosomes. The correlation matrix of these markers are:

$$\text{Cor}_{\text{chr1}=} \begin{pmatrix} 1 & 0.254 & 0.056 \\ & 1 & -0.329^{*} \\ & & 1 \end{pmatrix}, \quad \text{Cor}_{\text{chr3}=} \begin{pmatrix} 1 & 0.75^{**} & 0.110 \\ & 1 & 0.088 \\ & & 1 \end{pmatrix}$$

$$\mathrm{Cor}_{\mathrm{chr4}=}\begin{pmatrix} 1 & -0.056 & -0.155 & 0.021 \\ & 1 & -0.133 & 0.43^{*} \\ & & 1 & -0.069 \\ & & & 1 \end{pmatrix},$$

$$\mathrm{Cor}_{\mathrm{chr11}=}\begin{pmatrix} 1 & -0.076 & -0.099 \\ & 1 & 0.251 \\ & & 1 \end{pmatrix}, \quad \mathrm{Cor}_{\mathrm{chr13}=}\begin{pmatrix} 1 & 0.428^{*} & 0.238 \\ & 1 & 0.231 \\ & & 1 \end{pmatrix}$$

$$\mathrm{Cor}_{\mathrm{chr25}=}\begin{pmatrix} 1 & 0.009 & -0.001 & -0.111 \\ & 1 & -0.045 & 0.05 \\ & & 1 & 0.108 \\ & & & 1 \end{pmatrix},$$

$$\mathrm{Cor}_{\mathrm{chr28}=}\begin{pmatrix} 1 & -0.244 & 0.007 & -0.130 \\ & 1 & 0.235 & 0.081 \\ & & 1 & 0.079 \\ & & & 1 \end{pmatrix},$$

$$\mathrm{Cor}_{\mathrm{chr6}}\left(\mathrm{SNP}_{293}, \mathrm{SNP}_{314}\right) = 0.386^{*},$$

$$\mathrm{Cor}_{\mathrm{chr8}}\left(\mathrm{SNP}_{371}, \mathrm{SNP}_{387}\right) = 0.005,$$

$$\mathrm{Cor}_{\mathrm{chr12}}(\mathrm{SNP}_{571}, \mathrm{SNP}_{578}) = 0.015\ ,$$

$$\mathrm{Cor}_{\mathrm{chr18}}(\mathrm{SNP}_{803}, \mathrm{SNP}_{813}) = 0.313^{*},$$

$\text{Cor}_{\text{chr22}}(\text{SNP}_{950}, \text{SNP}_{951}) = 0.404^{*}$.

where star denotes significant correlation. Clearly, detected SNPs in most of the Chromosomes are weekly correlated and we can be sure that they are not detected due to the LD. Although, the correlation matrix of chromosome 3 shows two markers closely link to each other. Hence, they might be identified due to the high correlation.

6.6.2 Epistatic Model

In this section, we applied the epistatic model introduced in Chapter 5. As we have seen in last section, the impact of markers on the trait are small. Hence, the single marker analysis does not have power to detect true association. Instead of selecting based on threshold 10^5 as typical threshold in single marker analysis, we considered the selected SNPs in the previous section. The model thereby includes the selected SNPs in previous section and their pairwise interactions.

To select hyperparameters of the model that affected on the rate of shrinkage, c_1 and c_2 in (5.3), we ran a 10-fold cross validation. Table 6.6 represents the average of 10 out-of-samples MSPEs for different values of hyperparameters and the standard deviation from 100 bootstrap samples in subscript.

Table 6.6: Average of 10 out-of-samples MSPEs for different values of hyperparameters and their standard deviations from 100 bootstrap samples in subscript.

c_2 \ c_1	2	2.5	3	3.5
2.5	$0.556_{(0.036)}$	-	-	-
3.5	$0.526_{(0.033)}$	$0.538_{(0.031)}$	$0.555_{(0.031)}$	-
4.0	$0.516_{(0.026)}$	$0.509_{(0.025)}$	$0.520_{(0.027)}$	-
4.5	$0.550_{(0.032)}$	$0.547_{(0.32)}$	$0.549_{(0.029)}$	$0.550_{(0.031)}$

As can be seen from table 6.6, values 2.5 and 4 are the best choice for c_1

and c_2 respectively. By these choices of hyperparameters, we ran the MCMC for the whole dataset. Then similar to our procedure in last section we selected the SNPs with marginal effects or epistatic effects on the traits. The Table 6.7 gives the estimated effect sizes and the marginal heritabilities. The total genetic variance contributed by the main and epistatic effects of the markers was 0.517 when the total phenotypic variance was estimated 1.723. Therefore, the total heritability for this model is larger than previous model with no epistatic effect.

Table 6.7: The estimated marginal and epistatic effects with total heritability above 0.2.

index(i, j)	$\hat{\beta}$	$\hat{h}^2(\%)$
(40 , 1032)	0.0661	0.7848
(54 , 54)	0.0678	0.7829
(54 , 813)	-0.0994	0.9676
(122 , 332)	-0.0688	0.7770
(122 , 620)	-0.0695	0.7714
(122 , 1111)	-0.0654	0.7624
(151 , 620)	-0.0681	0.7805
(180 , 180)	-0.0704	0.8247
(192 , 420)	0.0651	0.7507
(192 , 770)	0.0767	0.7811
(293 , 631)	0.1033	1.1427
(332 , 571)	-0.0737	0.8079
(371 , 631)	0.0726	0.8044
(387 , 420)	0.1104	1.4580
(420 , 1032)	0.0674	0.7730
(420 , 1111)	0.0685	0.7676
(571 , 1103)	-0.0704	0.7799
(624 , 1114)	0.0719	0.8010
(631 , 652)	-0.0690	0.7549
(770 , 770)	0.0671	0.7743
(770 , 1111)	0.0761	0.8244
(813 , 813)	-0.0660	0.8720
(950 , 950)	0.0687	0.7921
(1032, 1032)	0.0869	0.7719
(1039, 1039)	-0.0764	0.7514
(1106, 1106)	0.0811	0.8485

6.7 Discussion

In genome-wide association studies, preliminary analysis plays crucial role to prevent spurious association. To identify true association, we need to account for population structures like population stratification and related samples which are very common in genetic problems. Although, we can improve the power of detection with preliminary analysis, identifying associated SNPs

is very difficult in some problem that the heritability of the trait is low. In these kinds of problems, single marker analysis may not be powerful to detect any association. Therefore, the multi-stage approach can be more appropriates. Our proposed model that is based on multi-stage analysis shows a good predictive performance for problems with weak marker effects in compare to the BL and BR. It also reveals that in genetic problems, we face with a complex network and considering epistatic effects can capture more heritability. This is an issue that cannot be characterized in single marker analysis.

Bibliography

Achlioptas, D. (2003). Data-friendly random projections. *Journal of Computer and System Sciences* **66**, 671 – 87.

Armagan, A., Dunson, D. B. & Lee, J. (2013). Generalized double pareto shrinkage. *Statistica Sinica* **23**, 119 – 143.

Bae, K. & Mallick, B. (2004). Gene selection using a two-level hierarchical Bayesian. *Bioinformatics* **20**, 3423 – 30.

Ball, R. D. (2011). Experimental designs for robust detection of effects in genome-wide case-control studies. *Genetics* **189**, 1497 — 514.

Benjamini, Y. & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. B* **85**, 289 – 300.

Benjamini, Y. & Yekutieli, D. (2001). The cotrol of the false discovery rate in multiple testing under dependency. *Ann. Statist* **29**, 1165 — 88.

Carlin, B. P. & Louis, T. A. (2001). *Bayes and empirical Bayes methods for data analysis*. Secod edition. London/ Chapman and Hall.

Carvalho, C., Polson, N. & Scott, J. (2009). Handling sparsity via the horseshoe. *Journal of Machine Learning Research* **5**, 73 – 80.

Carvalho, C. M., Polson, G. N. & Scott, G. J. (2010). The horseshoe estimator for sparse signals. *Biometrika* **97**, 465 – 80.

Castillo, I. & van der Vaart, A. (2012). Needles and straws in a haystack: Posterior concentration for possibly sparse sequences. *Journal of Machine Learning Research* **40**, 2069 – 2101.

Cui, W. & George, E. I. (2008). Empirical Bayes vs. fully Bayes variable selection. *J. Statist. Plann. Inference* **138**, 888 – 900.

Dasgupta, S. & Gupta, A. (2003). An elementary proof of a theorem of johnson and lindenstrauss. *Random Structures and Algorithms* **22**, 60 – 65.

David, A. P. (1973). Posterior expectations for large observations. *Biometrika* **60**, 664 – 67.

Diabetes Genetics Initiative of Broad Institute of Harvard and MIT (2007). Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. *Science* **316**, 1331 — 36.

Diebolt, J. & Robert, C. P. (1994). Estimation of finite mixture distributions through Bayesian sampling. *J. Roy. Statist. Soc. B* **56**, 363 – 75.

Dudbridge, F. & Gusnato, A. (2008). Estimation of significance threshold for genomewide association scans. *Genet. Epidemiol.* **32**, 227 – 34.

Efron, B. (2004). Large-scale simultaneous hypothesis testing: The choice of a null. *J. Amer. Statist. Assoc.* **99**, 96 — 104.

Efron, B. (2007). Correlation and large-scale simultaneous significance testing. *J. Amer. Statist. Assoc.* **102**, 93 — 103.

Efron, B., Hastie, T., Johnstone, I. & Tibshirani, R. (2004). Least angel regression. *Ann. Statist.* **32**, 407 – 499.

Efron, B. & Moris, C. (1975). Data analysis using stein's estimator and its generalizations. *J. Amer. Statist. Assoc.* **70**, 311 – 19.

Efron, B., Tibshirani, R., Storey, J. D. & Tusher, V. (2001). Empirical Bayes analysis of a microarray experiment. *Ann. Statist.* **96**, 1151 – 60.

Erkanli, A. (1994). Laplace approximation for posterior expectations when the mode occurs at the boundary of the parameter space. *J. Amer. Statist. Assoc.* **89**, 250 – 58.

Evans, D. M., Marchini, J., Morris, A. P. & Cardn, L. R. (2006). Two-stage two-locus models in genome-wide association. *PLoS Genet.* **2**. E157. DOI: 10.1371/journal.pgen.0020157.

Fan, J. & Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96**, 1348 – 60.

Fan, J. & Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space (with discussion). *J. R. Stat. Soc. B* **70**, 849 – 911.

Fard, M. M., Grinberg, Y., Pineau, J. & Precup, D. (2012). Compressed least-squares regression on sparce space. *Proc. of the Twenty-Sixth AAAI Conference on Artificial Intelligence* .

Fernandez, C., Ley, E. & Steel, M. F. (2001). Benchmark priors for Bayesian model averaging. *J. Econometrics* **100**, 381 – 427.

Figueiredo, M. (2003). Adaptive spareseness for supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **25**, 1150 – 9.

Foster, D. P. & George, E. I. (1994). The risk inflation criterion for multiple regression. *Ann. Statist.* **22**, 1947 – 75.

Frühwirth-Schnatter, S. & Wagner, H. (2010). Bayesian variable selection for random intercept modeling of gaussian and non-gaussian data. *Bayesian Statistics 9. In Bernardo, J., Bayarri, M., Berger, J. O., Dawid, A. P., Heckerman, D., Smith, A. F. M., West, M., eds* .

Gelman, A. (2003). Bugs.R: function for calling Bugs from R. www.stat.columbia.edu/ gelman/bugsr/ .

George, E. I. & Foster, D. P. (2000). Calibration and empirical Bayes variable selection. *Biometrika* **87**, 731 – 47.

George, E. I. & McCulloch, R. E. (1993). Variable selection via Gibbs sampling. *J. Amer. Statist. Assoc* **881 – 9**.

George, E. I. & McCulloch, R. E. (1997). Approaches for Bayesian variable selection. *Statistic. Sinica* **7**, 339 – 73.

Goemana, J. J. & Aldo Solarib, A. (2014). Tutorial in biostatistics: multiple hypothesis testing in genomics. *Statist. Med.* Doi: 10.1002/sim.6082.

Griffin, J. E. & Brown, P. J. (2007). Bayesian adaptive lasso with non-convex penalization. *Technical Report* .

Griffin, J. E. & Brown, P. J. (2010). Inference with normal-gamma prior distribution in regression problem. *Bayesian Analysis* **5**, 171 – 88.

Guan, Y. & Stephens, M. (2011). Bayesian variable selection regression for genome-wide association studies, and other large-scale problems. *Ann. App. Statist.* **5**, 1780 – 815.

Guhaniyogi, R. & Dunson, D. B. (2013). Bayesian compressed regression. *Arxiv Preprint arxiv* **1303.0642v2**.

Hans, C. M. (2009). Bayesian lasso regression. *Biometrika* **96**, 835 – 45.

Hasio, C. K. (1997). Approximate Bayes factors when a mode occurs on the boundary. *J. Amer. Statist. Assoc.* **92**, 656 – 63.

Hoh, J., Wille, A., Zee, R., Cheng, S., Reynolds, R., Lindpaintner, K. & Ott, J. (2000). Selection SNPs in two-stage analysis of disease association data: a model-free approach. *Ann. Hum. Genet.* **64**, 413 – 17.

Hoti, F. & Sillanpaa, M. J. (2006). Bayesian mapping of genotype × expression interactions in quantitative and qualitative traits. *Heredity* **97**, 4 – 18.

Indyk, P. & Motwani, R. (1998). Approximate nearest neighbors: Towards removing the curse of dimensionality. *Proc. of STOC.* , 604 – 13.

Jeffreys, H. (1961). *Theory of Probability.* Oxford Univ. Press.

Johnson, W. B. & Lindenstrauss, J. (1984). Extensions of lipschitz mapping into hilbert space. *Contemporary Mathematics* **26**, 189 – 206.

Kass, R. E. & Raftery, A. E. (1995). Bayes factors. *J. Amer. Statist. Assoc.* **90**, 773 – 95.

Lee, M. L. T., Kuo, F., Whitmore, G. & Sklar, J. (2000). Importance of replication in microarray gene expression studies: Statistical methods and evidence from repetitive cDNA hybridizations. *Proc. Nat. Acad. Sci.* **97**, 9834 – 38.

Lee, S., Epstein, M., Duncan, R. & Lin, X. (2012). Sparse principal component analysis for identifying ancestry-informative markers in genome-wide association studies. *Genet. Epidemiol.* **36**, 293 – 302.

Ley, E. & Steel, M., F. J. (2007). Jointness in Bayesian variable selection with applications to growth regression. *Journal of Macroeconomics* **29**, 476 —- 93.

Li, J., Das, K., Fu, G., Li, R. & Wu, R. (2011). The Bayesian lasso for genome-wide association studies. *Bioinformatics* **27**, 516 – 23.

Li, P., Hastie, J. T. & Church, W. K. (2006). Very sparse random projections. *Proc. of KDD, Philadelphia* , 278 – 96.

Liang, F., Paulo, R., Molina, G., Clyde, M. A. & Berger, J. O. (2008). Mixture of g-priors for Bayesian variable selection. *J. Amer. Statist. Assoc.* **103**, 410 – 23.

Liu, C., Rubin, D. B. & Wu, Y. N. (1998). Parameter expansion to accelerate EM: the PX-EM algorithm. *Biometrika* **85**, 755 – 70.

Maillard, O. & Munos, R. (2009). Compressed least-squares regression. *Proc. of Advances in neural information processing systems* .

Maruyama, Y. & George, I. E. (2011). Fully Bayes factors with a generalized g-prior. *Ann. Statist.* **39**, 2740 – 65.

Masoeliez, C. (1975). Approximation non-gaussian filtering with linear state and observation relations. *IEEE. Trans. Auto. Contr.* **20**, 107 –10.

Murcray, C. E., Lewinger, J. P. & Gauderman, W. J. (2009). Gene-environment interaction in genome-wide association studies. *Am. J. Epidemiol.* **169**, 219 – 26.

Newton, M., Kendziorski, C., Richmond, C., Blattner, F. & Tsui, K. (2001). On differential variability of expression ratios: Improving statistical inference about gene expression changes from microarray data. *J. Comp. Biology* **8**, 37 – 52.

Park, T. & Casella, G. (2008). The Bayesian lasso. *J. Amer. Statist. Assoc.* **103**, 681 – 86.

Paul, D., Bair, E., Hastie, T. & Tibshirani, R. (2008). Preconditioning for feature selection and regression in high-dimensional problems. *Ann. Stat.* **36**, 1595 – 618.

Pericchi, L. R. & Smith, A. (1992). Exact and approximate posterior moments for a normal location parameter. *J. R. Statist. Soc. B* **54**, 793 – 804.

Polson, N. G. (1991). A representation of the posterior mean for a location model. *Biometrika* **78**, 426 – 30.

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

J. M., Bayarri, M. J., Berger, J. O., Dawid, A. P., Heckerman, D., Smith, A. F. M., and West, M. eds. , 501 – 38.

Redner, R. A. & Walker, H. F. (1984). Mixture densities, maximum likelihood and the EM algorithm. *SIAM Rev.* **26**, 195 – 239.

Scott, G. J. & Berger, O. J. (2010). Bayesian and empirical-Bayes multiplicity adjustment in the variable-selection problem. *Ann. Statist.* **38**, 2587 – 619.

Smith, M. & Kohn, R. (1996). Nonparametric regression using Bayesian variable selection. *J. Econometrics* **75**, 317 – 43.

Stein, C. (1956). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. *Proc. of the Third Berkeley Symposium on Mathematical Statistics and Probability* **1**, 197 – 206.

Stephens, M. & Balding, D. J. (2009). Bayesian statistical methods for genetic association studies. *Nat. Rev. Genet.* **10**, 681 – 690.

Storey, J. (2002). A direct approach to false discovery rates. *J. Roy. Statist. Soc. B* **64**, 479 – 98.

Storey, J. D. (2003). The positive false discovery rate: a Bayesian interpretation and the q-value. *Ann. Statist.* **31**, 2013 — 35.

Storey, J. D. & Tibshirani, R. (2003). Statistical significance for genomewide studies. *PNAS* **100**, 9440 – 45.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. Royal. Statist. Soc.* **58**, 267 – 88.

Vempala, S. (2004). *The random projection method*. American Mathematical Society, Province, RI.

Wakefield, J. (2007). A Bayesian measure of the probability of false discovery in genetic epidemiology studies. *Am. J. Hum. Genet.* **81**, 208 — 27.

Wellcome Trust Case Control Consortium (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661 — 78.

Yi, N. & Xu, S. (2008). Bayesian lasso for quantitative trait loci mapping. *Genetics* **179**, 1045 – 55.

Zellner, A. (1986). On assessing prior distributions and Bayesian regression analysis with g-prior distributions. *In Bayesian inference and decision techniques: Essays in honor of Bruno de Finetti* , 233 – 43, North-Holland/ Elsevier.

Zellner, A. & Siow, A. (1980). Posterior odds ratios for selected regression hypotheses. *Proc. of the First International Meeting, Valencia* , 585 – 603.

Zhang, C. H. (2010). Nearly unbiased variable selection under minimax concave penalty. *Ann. Statist.* **38**, 894 – 942.

Zhao, P. & Yu, B. (2006). On model selection consistency of lasso. *Journal of Machine Learning Research* **7**, 2541 – 63.

Zheng, G., Song, K., & Elston, R. C. (2007). Adaptive two-stage analysis of genetic association in case-control designs. *Hum. Hered.* **63**, 175 – 86.

Zou, H. (2006). The adaptive lasso and its oracle properties. *J. Amer. Statist. Assoc.* **101**, 1418 – 29.

Zou, H. & Hastie, T. (2005). Regularization and variable selection via the elastic net. *J. R. Statist. Soc. B* **67**, 301 – 320.

Zou, H., Hastie, T. & Tibshirani, R. (2006). Sparse principal component analysis. *Journal of Computational and Graphical Statistics* **15**, 265 — 286.

Akram Yazdani

CURRICULUM VITAE

Contact Information

University of Padova
Department of Statistics
via Cesare Battisti, 241-243
35121 Padova. Italy.

e-mail: yazdani@stat.unipd.it
ak_yazdani@yahoo.com

PhD student

Since January 2011;
PhD Student in Statistical Sciences, University of Padova.
Thesis title: Statistical Approaches in Genome-Wide Studies
Supervisor: Prof. Monica Chiogna

Research interests

- *Statistical methods for high dimensional data: Classification and clustering, Bayesian and computational statistics, compressive methods for massive datasets, dimensionality reduction, hypothesis testing. The motivation comes from applications in genetics and neurosciences.*

Education

September 2002 – December 2005
Master degree in Statistics .
University of Shahid Beheshti, Department of Mathematics, Tehran, Iran
Title of dissertation: Estimation the Poisson Regression Coefficients by Maximum Likelihood, and Hierarchical Bayes methods.
Supervisor: Prof. Rahim Shahlaei
Final mark: A^+

September 1998 – September 2002
Bachelor degree in Statistics.
University of Allameh Tabatabaei, Department of Economics, Tehran, Iran
Title of dissertation: Analysis and design with aim of automating system
Supervisor: Prof. Mohammad Reza Zandinia
Final mark: A^+

Further education

November 2012
Genetic Mapping Course
Organizer: Baylor College of Medicine, Houston, Texas, USA and Max-delbrück Centrum Für Molekulare Medizin, Berlin, Germany
Instructor: Prof. Suzanne M. Leal and Micheal Nothnagel

January 2013 – February 2013
Winter School in genetics
Organizer: Catholic University
Instructor: Prof. Paolo Ajmone and Prof. Alessio Valentini

Work experience

September 2004 – Jun 2010
Payam e Noor Univesity, Iran
Lecturer

March 2012 – February 2013
Catholic University in Milan, Italy
Researcher

Awards and Scholarship

1998
B.Sc. scholarship (Iranian Government Grant)

2002
M.Sc. scholarship (Iranian Government Gran)

2012
Postdoctoral fellowship (Catholic University)

Computer skills

- *Operating System: Linux, Windows*
- *Programming: C++, C, R, SQL, Matlab*
- *Software: WinBUGS, SPSS, Plink*
- *Markup Language: LaTex*

Language skills

Persian: native; English: fluent (written/spoken);

Teaching experience

February 2013 – February 2013
Course name: Introduction to Bayesian Statistics
Degree: Graduate student
Teaching task; Lecturer, 9 hours
Institution: Catholic University in Piachenza, Italy
Instructor: Prof. Paolo Ajmone Marsan

September 2004 – Jun 2010
Course names: Matematical Statistics, Probability, Experimental Design
Degree: Under Graduate
Teaching task: Lecturer
Institution: Payam e Noor University