A Bayesian Group Sparse Multi-Task Regression Model for Imaging Genomics

by

Keelin Greenlaw

B.Sc., Brock University, 2011

A Thesis Submitted in Partial Fulfilment of the

Requirements for the Degree of

MASTER OF SCIENCE

in the Department of Mathematics and Statistics

by

Keelin Greenlaw

B.Sc., Brock University, 2011

Supervisory Committee

Dr. Mary Lesperance, Co-supervisor

(Department of Mathematics and Statistics)

Dr. Farouk Nathoo, Co-supervisor

(Department of Mathematics and Statistics)

Dr. Mary Lesperance, Co-supervisor

(Department of Mathematics and Statistics)

Dr. Farouk Nathoo, Co-supervisor

(Department of Mathematics and Statistics)

ABSTRACT

Recent advances in technology for brain imaging and high-throughput genotyping have motivated studies examining the influence of genetic variation on brain structure. In this setting, high-dimensional regression for multi-SNP association analysis is challenging as the brain imaging phenotypes are multivariate and there is a desire to incorporate a biological group structure among SNPs based on their belonging genes. Wang et al. (Bioinformatics, 2012) have recently developed an approach for simultaneous estimation and SNP selection based on penalized regression with regularization based on a novel group $\ell_{2,1}-norm$ penalty, which encourages sparsity at the gene level. A problem with the proposed approach is that it only provides a point estimate. We solve this problem by developing a corresponding Bayesian formulation based on a three-level hierarchical model that allows for full posterior inference using Gibbs sampling. For the selection of tuning parameters, we consider techniques based on: (i) a fully Bayes approach with hyperpriors, (ii) empirical Bayes with implementation based on a Monte Carlo EM algorithm, and (iii) cross-validation (CV).

When the number of SNPs is greater than the number of observations we find that both the fully Bayes and empirical Bayes approaches overestimate the tuning parameters, leading to overshrinkage of regression coefficients. To understand this problem we derive an approximation to the marginal likelihood and investigate its shape under different settings. Our investigation sheds some light on the problem and suggests the use of cross-validation or its approximation with WAIC (Watanabe, 2010) when the number of SNPs is relatively large. Properties of our Gibbs-WAIC approach are investigated using a simulation study and we apply the methodology to a large dataset collected as part of the Alzheimer's Disease Neuroimaging Initiative.

Contents

List of Tables

List of Figures

ACKNOWLEDGEMENTS

I will never be able to thank my supervisors, Dr. Mary Lesperance and Dr. Farouk Nathoo, enough for the guidance and support they have given me over the course of this project. Their talent, expertise, and patience seem only to be surpassed by an enthusiasm for research, an enthusiasm that is both inclusive and contagious. Before this past year, I'm entirely sure that *cool* and *fun* would not have been the descriptive words I would have chosen for *Statistics*, but working with Mary and Farouk has completely changed my opinion on this. Thank you.

I would also like to thank Elena Szefer and Dr. Jinko Graham for providing the processed data used in this project, as well as the source of the data, the Alzheimer's Disease Neuroimaging Initiative. Thank you to Dr. Belaid Moa, who showed incredible patience in teaching me to utilise the resources provided by Compute Canada, and to Compute Canada for the use of their clusters.

To my parents, who have been supportive and understanding throughout all the difficult moments of learning and growing, thank you. Thank you to my brother, who has likely been the person who has challenged me the most through it all. I feel grateful to be able to call him a friend.

Finally, thank you to Chris for always finding time to laugh with me. This has made all the difference.

DEDICATION

To my grandfather, who valued education above all else.

To my grandmother, who, at 90 years old, continues to teach me the importance of long walks, good company, and ice cream.

Chapter 1

Introduction

1.1 Background

One focus of genomic research aims at finding genetic variations among people and using these small differences in genetic material to predict a person's risk of particular diseases (National Library of Medicine (US), 2015). Identifying these genetic markers may lead to an earlier diagnosis, and hopefully earlier treatment and improved prognosis.

DNA, the carrier of genetic information, is a code made up of four chemical bases: adenine (A), guanine (G), cytosine (C), and thymine (T). The bases pair up with each other to form base pairs; A always pairs with T, and G always pairs with C (McEntyre J, Ostell J, 2015). The human genome consists of about 3 billion base pairs (National Institutes of Health, 2015). Each base is also attached to a sugar molecule and a phosphate molecule. Together, a base, sugar, and phosphate make up a nucleotide. A section of DNA is illustrated in Figure 1.1.

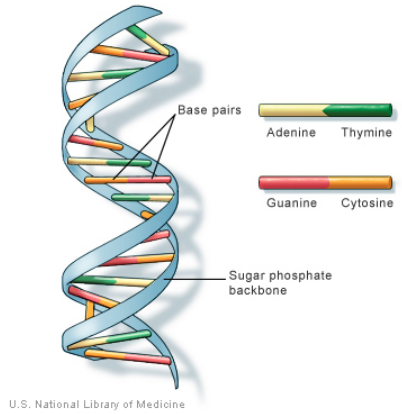


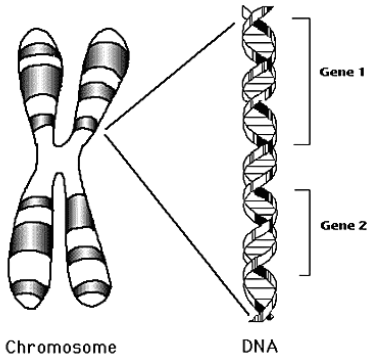
Figure 1.1: DNA is formed by base pairs attached to a sugar-phosphate backbone.

DNA contains the instructions needed for an organism to develop, survive and reproduce (National Library of Medicine (US), 2015). Genes are the basic physical and functional units of genetic material and are represented as sequences of nucleotides, e.g. AACTA (Figure 1.2a). Most gene sequences are conserved across all humans, however, there are locations on DNA where variability exists.

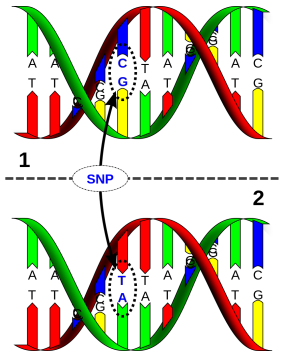
Single nucleotide polymorphisms (SNPs) are the most common type of genetic variation among people. A SNP is a single nucleotide location where human DNA varies (Figure 1.2b); there are approximately 10 million SNPs in the human genome (National Library of Medicine (US), 2015).

A chromosome is an organised package of DNA found within the nucleus of each cell (National Institutes of Health, 2015). In humans, cells other than human sex cells are diploid. A diploid cell has paired chromosomes, one from each parent (National Institutes of Health, 2015). Alleles are the possible alternative nucleotides at a specific position in the sequence for a gene (McEntyre J, Ostell J, 2015); the minor allele refers to the allele that occurs less frequently in the population. Most human cells are diploid, and thus contain two copies of each allele. SNP data for an individual at a given SNP is given by $0, 1, 2$, indicating

that their DNA contains zero, one, or two copies of the minor allele.



(a) Genes are composed of DNA sequences.



(b) Single nucleotide polymorphism

Figure 1.2: Genetic terms

Genetic variations are the focus of genome-wide association studies where the goal may be to identify SNPs that occur more frequently in individuals with a particular disease compared to those without the disease. This is an example of a case-control study where the phenotype (an individual's observable trait) is the disease status.

1.2 Imaging genetics

Quantitative imaging-derived traits are thought to have more direct links to genetic variations than diagnostic measures based on cognitive or clinical assessments (Ge, Feng, Hibar, Thompson, & Nichols, 2012). Thus, these quantitative traits (QTs) might offer increased statistical power to detect the association between specific genes or SNPs and various brain diseases (Ge et al., 2012). Imaging genetics is an emergent research field where these associations, between SNPs and imaging phenotypes as QTs, are evaluated. This has driven the collection of large amounts of imaging and genetic data, such as in the Alzheimer's Disease Neuroimaging Initiative (ADNI).

The nature of these data and underlying biological mechanisms presents many statistical challenges. Case-control studies only look at one binary outcome, whereas the responses for QTs are multivariate continuous data. Pairwise univariate analysis was commonly used in traditional association studies, but this approach treats SNPs and QTs as independent units (Wang et al., 2012). Functionality of the human brain regularly involves more than one cerebral component (Wang et al., 2012) and therefore we wish to model this dependence to avoid losing the underlying interacting relationships. Likewise, multiple SNPs from one gene often jointly carry out genetic functionalities (Wang et al., 2012). Furthermore, the data is high-dimensional and it is suspected that only a small number of SNPs are risk factors for disease and associated with intermediate imaging traits. Therefore, in addition to incorporating interacting relationships, there is also the desire to select a subset of important SNPs from the larger group of SNPs under investigation. Some potential brain-wide, genome-wide association studies are based on penalised and sparse regression techniques, including ℓ_1 regularisation in the *least absolute shrinkage and selection operator* (lasso) (Tibshirani, 1996). These methods are able to localise brain regions and genomic regions from high dimensional data, and have been reported to show increased statistical power (Ge et al., 2012).

1.3 Objective

We focus on multivariate phenotypes (volumetric and cortical thickness values) of moderate dimension (e.g. $10 - 30$) derived from MRI for certain regions of interest (ROIs). Figure 1.3 gives an example of how ROIs might be defined; a brain summary measure is provided for each sub-region. The genetic data, which we wish to relate to imaging phenotypes, are a set of SNPs chosen from a specific set of genes.

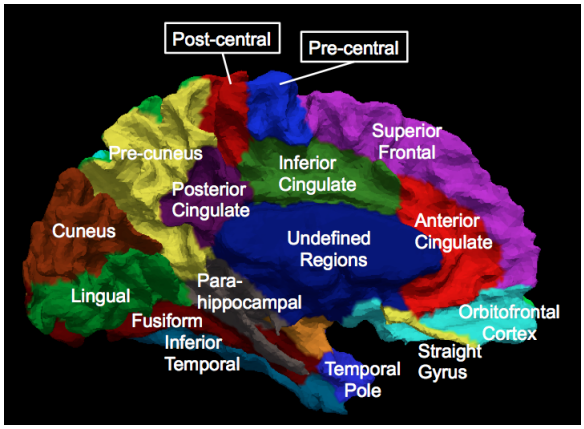


Figure 1.3: The medial surface of the human cerebral cortex parcellated into 24 sub-regions.

With the various statistical challenges in mind, we develop a Bayesian multivariate regression approach based on a continuous shrinkage prior that encourages sparsity and induces dependence in regression coefficients corresponding to SNPs within the same gene, and across components of imaging phenotypes. Though our approach is related to the Bayesian group lasso (Kyung, Gill, Ghosh, & Casella, 2010; Park & Casella, 2008) adapted for multivariate phenotypes, it is primarily motivated by Wang et al. (2012), who recently proposed a new framework, *Group-sparse multi-task regression and feature selection* (G-SMuRFS) for identifying quantitative trait loci. Built upon multivariate regression analysis, Wang et al. (2012) take into account the interrelated structure within and between SNPs and QTs by using a new form of regularisation, group $\ell_{2,1} - norm$, for group structure sparsity, in addition to $\ell_{2,1} - norm$ regularisation for individual structured sparsity. Their method, however, only provides point estimates of regression coefficients. In order to obtain valid measures of variability, we develop an equivalent hierarchical Bayesian model, which allows for inference based on posterior distributions. These measures of variability are the principal motivations and contributions of this body of work.

1.4 Outline

The remainder of this thesis is structured as follows: Chapter 2 gives a brief description of the data used in this project (for more information on data sources and preprocessing steps see Szefer (2014)). Chapter 3 reviews some of the common themes employed throughout this thesis. We begin by reviewing some basic concepts and methods used in Bayesian statistics and proceed to give a quick overview the lasso (Tibshirani, 1996) and some lasso variations (Tibshirani, Saunders, Rosset, Zhu, & Knight, 2005; Yuan & Lin, 2006; Zou & Hastie, 2005). We then discuss how these sparse regression methods can be modelled in a Bayesian framework. In Chapter 4, the estimator proposed by Wang et al. (2012) is presented in greater detail. This chapter also sets up notation for data used in subsequent chapters. We develop our hierarchical Bayesian model in Chapter 5. We prove its correspondence to the Wang et al. estimator and show derivations of full conditional distributions for Gibbs sampling. Chapter 6 is dedicated to the discussion of selection of tuning parameters. We show some preliminary results based on the two commonly discussed methods. As a consequence of the problems that arise here, we derive an approximation to the marginal likelihood and study its shape in different simulation settings. Lastly, we finalise our method as Gibbs Sampling with selection of tuning parameters based on WAIC (Watanabe, 2010). Experimental results are presented in Chapter 7, beginning with a large simulation study where we evaluate the performance of our method as well as compare its performance to the Wang et al. estimator. The second part of this chapter includes an application of both the Wang et al. method and our Gibbs-WAIC Bayesian method to a dataset collected as part of the Alzheimer's Disease Neuroimaging Initiative (ADNI). Finally, we conclude in Chapter 8 with a brief discussion and possible future directions.

Chapter 2

Materials and data sources

Data used in preparation of this thesis were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this thesis. A complete listing of ADNI investigators can be found at:
http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf.

The specific genetic and structural MRI data used in this study were obtained from the Alzheimer's Disease Neuroimaging Initiative 1 (ADNI-1) database. The data has been collected and processed to be similar to the data utilised by Wang et al. (2012). We include genetic and brain measurement data on 632 subjects from 3 categories of disease status. The dataset includes 179 subjects who are cognitively normal (CN), 309 in the late mild cognitive impairment (LMCI) stage, and 144 subjects with Alzheimer's disease (AD).

2.1 SNP genotyping and group information

Genetic data used in this study are queried from the most recent genome build, as of December 2014, from ADNI-1 genomic data. In their study, Wang et al. (2012) include only SNPs belonging to the top 40 Alzheimer's Disease (AD) candidate genes listed on the AlzGene database as of June 10, 2010. After performing standard quality control and imputation their dataset contains 1224 SNPs from 37 genes. We include all SNPs belonging to these 37 genes from the latest genome build (December 2014) from ADNI-1 genomic data. After standard quality control and imputation (see Szefer (2014)), 510 SNPs from 33 genes remain. After excluding SNPs with missing values, the final dataset used in this study includes 486 SNPs from 33 genes (Figure 2.1). Examples of SNP counts included in our dataset are shown in Figure 2.2. SNP counts are centered prior to fitting the model.

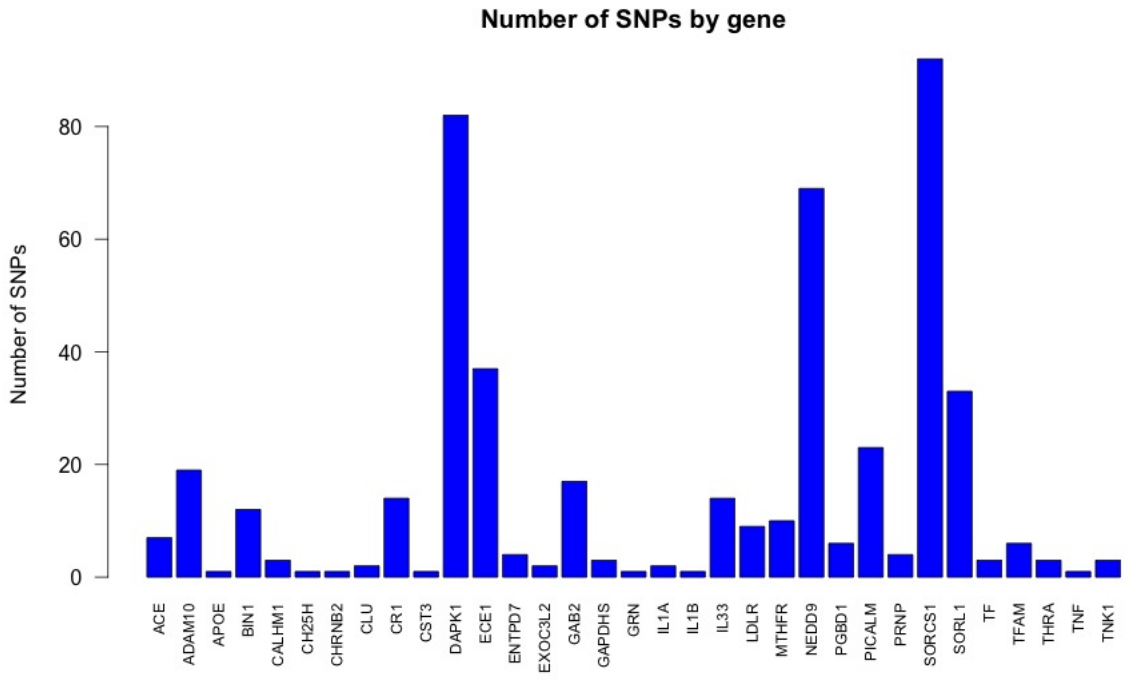


Figure 2.1: 33 AD risk factor genes used in this study and the number of SNPs in each.

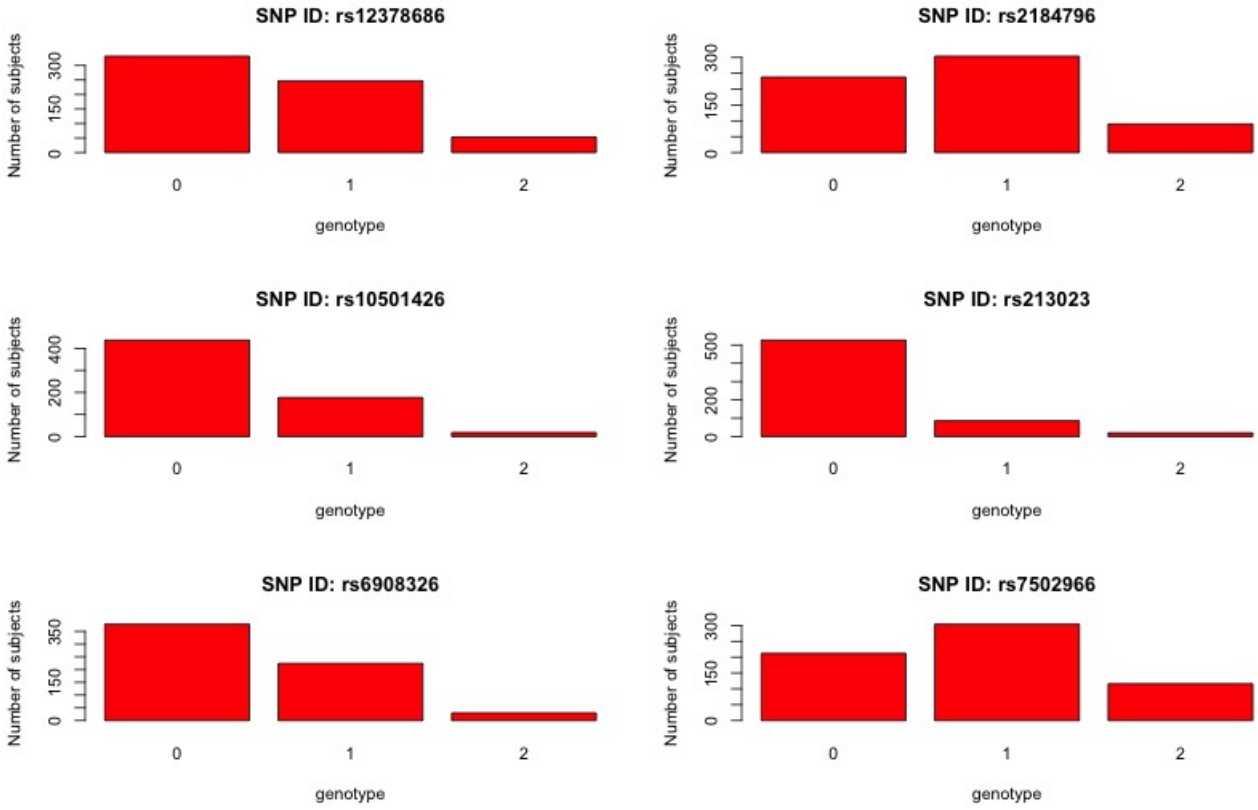


Figure 2.2: Example of SNP counts included in the dataset.

2.2 MRI analysis and extraction of imaging phenotypes

FreeSurfer is a brain imaging software package developed for analysing human brain magnetic resonance imaging (MRI) scan data (see: http://surfer.nmr.mgh.harvard.edu/). FreeSurfer is used to extract volumetric and cortical thickness values for regions of interest (ROIs), and to extract total intracranial volume (ICV) from scans of ADNI participants. A select 12 FreeSurfer measures from ROIs that are known to be related to Alzheimer's Disease (Wang et al., 2012) are included in this study. FreeSurfer measure IDs and ROI descriptions are shown in Table 2.1.

ID	Region of Interest (ROI)
Left_HippVol Right_HippVol	volume of hippocampus
Left_EntCtx Left_Parahipp Right_EntCtx Right_Parahipp	thickness of entorhinal cortex and thickness of parahippocampal gyrus
Left_Precuneus Right_Precuneus	thickness of precuneus
Left_MeanFront Right_MeanFront	mean thickness of caudal midfrontal, rostral midfrontal, superior frontal, lateral orbitofrontal, and medial orbitofrontal gyri and frontal pole
Left_MeanLatTemp Right_MeanLatTemp	Mean thickness of inferior temporal, middle temporal, and superior temporal gyri

Table 2.1: Volumetric/Thickness measures (FreeSurfer) and ROI descriptions

These measures are adjusted for age, gender, education, handedness and ICV based regression weights from healthy controls. The adjustments are made by fitting a linear model to those subjects who are cognitively normal (CN),

$$MRI\_measure \sim age + gender + education + handedness + ICV.$$

This fitted model is used to compute predicted values for the entire dataset. From the predicted values, residuals are obtained, which give the adjusted MRI measures. Additionally, MRI measures are centered and scaled prior to analysis.

Figures 2.3 and 2.4 show FreeSurfer measures from 4 ROIs before and after scaling and centering, respectively, with different colours indicating disease status of subjects. The figures demonstrate the relationship between disease status and brain summary measures from these ROIs; we see the group of CN subjects tending towards larger brain summary measures, and AD subjects towards smaller summary measures.

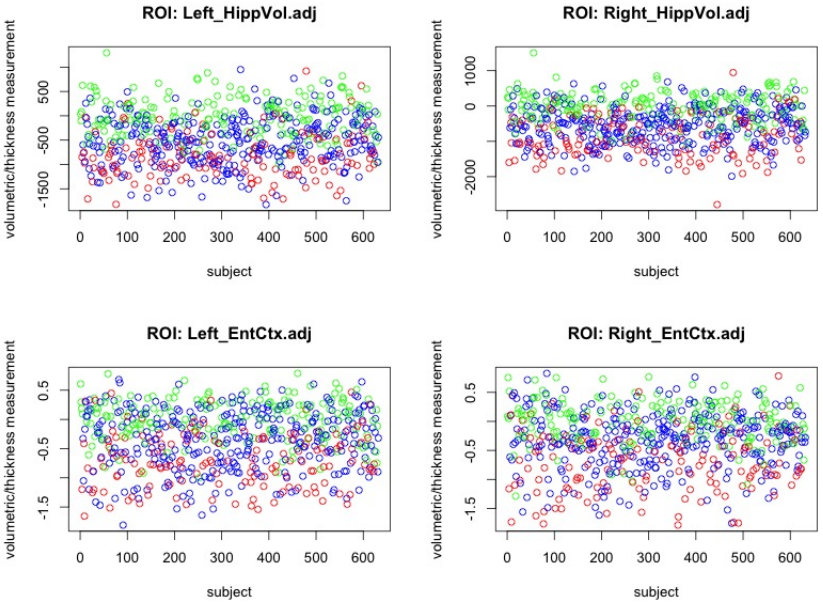


Figure 2.3: Adjusted FreeSurfer measurements. (Green = CN ; Blue = LMCI ; Red = AD)

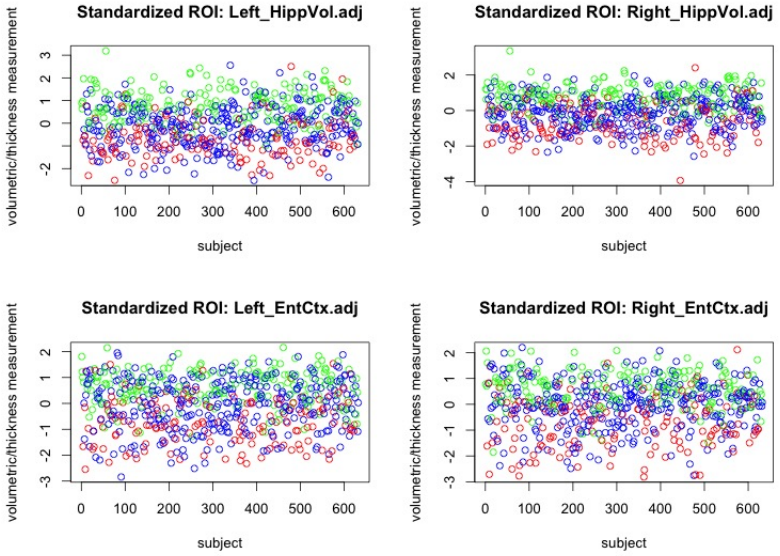


Figure 2.4: Adjusted FreeSurfer measurements after scaling and centering.

Chapter 3

Review:

Bayesian methods and the lasso

This chapter reviews relevant contributions of Kyung et al. (2010) to the field of lasso estimators. Section 3.1 briefly introduces some important definitions, concepts, and methods applied in the field of Bayesian statistics, which are referenced during the rest of the thesis. In Section 3.2 the lasso is introduced in more detail, and three lasso variations that arise from limitations to the original lasso are discussed. Section 3.3 demonstrates how the lasso can be modelled in a Bayesian framework and provides some insight into the necessary steps for Bayesian inference. Finally Section 3.4 discusses the advantages of the models described in this chapter.

3.1 Bayesian statistics and Gibbs sampling

3.1.1 Introduction to Bayesian statistics

Many statistical methods are based on frequentist inference where the unknown parameter(s) of interest, θ, is considered fixed and the data is considered a repeatable random sample. Bayesian inference, on the other hand, makes conclusions about θ in terms of probability statements (Gelman et al., 2013). These probability statements are conditional on the observed data, $\boldsymbol{y}$, which is regarded as a fixed realisation. It is at this fundamental level of conditioning on observed data that Bayesian inference differs from other approaches to statistical inference (Gelman et al., 2013).

Theorem 1. *Bayes' theorem*

$$p(\theta|\boldsymbol{y}) = \frac{p(\boldsymbol{y},\theta)}{p(\boldsymbol{y})} = \frac{p(\boldsymbol{y}|\theta)p(\theta)}{p(\boldsymbol{y})}$$

Theorem 1 is used to find an expression for the posterior distribution, $p(\theta|\boldsymbol{y})$, which represents the state of knowledge of θ given the data. The likelihood function is given by $p(\boldsymbol{y}|\theta)$ and $p(\theta)$ is the prior distribution of θ. The prior refers to the state of knowledge on θ before the data is collected. It is often convenient, for computation, if a conjugate prior is chosen.

Definition 1. *Conjugate prior*

Given a likelihood, the conjugate prior is the prior distribution such that the prior and posterior are in the same family of distributions. (Kulis, 2012)

Notice that the denominator in Theorem 1, $p(\boldsymbol{y})$, is the probability density of $\boldsymbol{y}$, which does not depend on θ. In the case where θ is continuous, we have the following expression

for the probability density of $\boldsymbol{y}$,

$$p(\boldsymbol{y}) = \int_{\theta} p(\boldsymbol{y}|\theta)p(\theta)d\theta.$$

This integral can be particularly difficult to solve, as it involves integrating over the entire parameter space of θ, which may be multidimensional. Luckily, when this is the case, we have methods based on Markov Chain Monte Carlo (MCMC) algorithms that allow the simulation of a large number of θ's from the posterior distribution. From these samples the posterior distribution can be approximated.

3.1.2 Gibbs sampling

MCMC methods are based on algorithms that follow properties of first-order Markov chains.

Definition 2. *First-order Markov chain*
A first-order Markov chain is defined as a sequence of random variables, $\theta^1, \theta^2, \ldots$ where, for any t, the distribution of θ^t given all previous θ's depends only on θ^{t-1}. (Gelman et al., 2013)

Gibbs sampling is a type of Markov chain algorithm that can be particularly useful when θ is high dimensional. The Gibbs sampler is defined in terms of subvectors of θ,

$$\theta = (\theta_1, \ldots, \theta_d).$$

Each iteration of the Gibbs sampler cycles through the subvectors of θ, drawing each subset conditional on all other parameters and the data, $\boldsymbol{y}$. Hence, there are d steps in each iteration t. At each iteration, each θ_j^t is sampled from the conditional distribution given all

other components of θ at their current values (Gelman et al., 2013),

$$P(\theta_j | \theta_{-j}^{t-1}, y) \;\; for \;\; j = 1, \ldots, d, \text{ where}$$

$$\theta_{-j}^{t-1} = (\theta_1^t, \ldots \theta_{j-1}^t, \theta_{j+1}^{t-1}, \ldots, \theta_d^{t-1}).$$

After what is known as the *burn-in* period, the samples of θ_j^t should have a distribution that converges to draws from a stationary distribution that approximates the target distribution, the full conditional posterior distribution of θ_j. From these large samples of θ_j we can compute estimators such as the mean and perform inference using quantiles of the large sample. Gibbs sampling, however, is only possible to implement when the full conditional distributions of $(\theta_1, \ldots, \theta_d)$ can be expressed in closed form.

3.2 The lasso and lasso variations

3.2.1 The lasso

The *Least Absolute Shrinkage and Selection Operator* (lasso) proposed by Tibshirani (1996) has the form

$$\hat{\boldsymbol{\beta}}_L = \arg\min_{\boldsymbol{\beta}} \left\{ (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) + \lambda \sum_{j=1}^{p} \left| \beta_j \right| \right\}.$$

The lasso minimises the residual sum of squares while penalising the sum of the absolute values of the coefficients. The penalty shrinks some coefficients and sets others to zero, which helps produce sparsity. This is a desirable property for variable selection. The tuning parameter, $\lambda \geq 0$, controls the amount of shrinkage and is typically estimated through cross-validation (Kyung et al., 2010). Numerous types of cross-validation methods exist, but

generally involve repeated partitioning of the data to training data, for parameter estimation, and test data, for parameter validation.

The lasso has demonstrated excellent performance in a variety of settings, but does exhibit some well-studied limitations (Kyung et al., 2010).

1. When considering the problem of selecting grouped variables, the lasso tends to select individual variables from the group.

2. When there exists multicollinearity among predictors, lasso prediction performance is dominated by ridge regression (another form of penalised regression).

3. If the predictors are ordered in some meaningful way, the ordering will be ignored by the lasso.

4. When $p > n$, the lasso cannot select more than n variables.

To compensate for limitations, several lasso variations have been developed. The next section introduces three lasso extensions.

3.2.2 Lasso variations

Group lasso

The group lasso for grouped variables, proposed by Yuan and Lin (2006), is defined as

$$\hat{\boldsymbol{\beta}}_G = \arg\min_{\boldsymbol{\beta}} \left\{ \left(\boldsymbol{y} - \sum_{\mathrm{k}=1}^{\mathrm{K}} \boldsymbol{X_k}\boldsymbol{\beta_k} \right)^T \left(\boldsymbol{y} - \sum_{\mathrm{k}=1}^{\mathrm{K}} \boldsymbol{X_k}\boldsymbol{\beta_k} \right) + \lambda \sum_{\mathrm{k}=1}^{\mathrm{K}} \left|\left|\boldsymbol{\beta_k}\right|\right|_{G_k} \right\},$$

where K is the number of groups, $\boldsymbol{\beta_k}$ is the vector of β_j's in group k, and $||\boldsymbol{\beta}||_G = (\boldsymbol{\beta}^T G \boldsymbol{\beta})^{\frac{1}{2}}$. In general $G_k = I_{m_k}$, where m_k is the number of coefficients in group k. Thus, the penalty term is a function of the sum over grouped coefficients, making the estimator useful in situations where the groups of the predictor variables are pre-determined.

Elastic net

$$\hat{\boldsymbol{\beta}}_{EN} = \arg\min_{\boldsymbol{\beta}} \left\{ (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) + \lambda_1 \sum_{j=1}^{p} |\beta_j| + \lambda_2 \sum_{j=2}^{p} |\beta_j|^2 \right\}$$

The elastic net was introduced by Zou and Hastie (2005) for an unknown group of variables and for predictors that present with multicollinearity. The penalty function of the elastic net includes the sum of both the absolute values and the squared values of the coefficients. The elastic net can be interpreted as a stabilized version of the lasso (Kyung et al., 2010), and naturally overcomes the difficulty of $p >> n$, while having the ability to perform grouped selection (Zou & Hastie, 2005).

Fused lasso

$$\hat{\boldsymbol{\beta}}_{F} = \arg\min_{\boldsymbol{\beta}} \left\{ (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) + \lambda_1 \sum_{j=1}^{p} |\beta_j| + \lambda_2 \sum_{j=2}^{p} \left|\beta_j - \beta_{j-1}\right| \right\}$$

The fused lasso was proposed specifically for problems where the predictors are ordered in a meaningful way (Tibshirani et al., 2005). An example of this meaningful ordering is gene expression data measured from a microarray, where highly correlated genes are near one another on the list because of their physical proximity. The fused lasso penalises both the coefficients, as well as the differences, which encourages sparsity of the coefficients and also their differences. Furthermore, Tibshirani et al. (2005) state that the fused lasso is highly effective in dealing with situations where $p >> n$.

3.3 Bayesian lasso hierarchical models

In this section an alternative approach for finding solutions to the lasso are described (Kyung et al., 2010).

3.3.1 The lasso in a Bayesian framework

Recall the lasso estimator:

$$\hat{\boldsymbol{\beta}}_L = \arg\min_{\boldsymbol{\beta}} \left\{ (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) + \lambda \sum_{j=1}^{p} \left|\beta_j\right| \right\}. \tag{3.1}$$

In a Bayesian framework we wish to not only find a point estimate for $\hat{\boldsymbol{\beta}}_L$, as in equation (3.1), but to base inference on the posterior distribution of $\boldsymbol{\beta}_L$. Theorem 1 gives the relationship $p(\boldsymbol{\beta}|\,\sigma^2, \boldsymbol{y}) \propto p(\boldsymbol{y}|\boldsymbol{\beta}, \sigma^2) p(\boldsymbol{\beta}|\,\sigma^2)$. The likelihood corresponds to the form found in basic linear regression,

$$\boldsymbol{y}\big|\boldsymbol{\beta}, \sigma^2 \sim N_n(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2\, I_n).$$

The penalty term included in the estimation of $\hat{\boldsymbol{\beta}}_L$ can be viewed as independent Laplace priors for each of the coefficients, β_j, and accordingly, the prior distribution of $\boldsymbol{\beta}$ takes the form

$$\pi(\boldsymbol{\beta}\big|\,\sigma^2) = \prod_{j=1}^{p} \frac{\gamma}{2\sigma} \exp\left\{ -\frac{\gamma}{\sigma} |\beta_j| \right\} \quad (p \textit{ independent Laplace priors}).$$

We now have the expression

$$p(\boldsymbol{\beta}|\,\sigma^2, \boldsymbol{y}) \propto \exp\left\{ -\frac{(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})}{2\,\sigma^2} - \frac{\gamma}{\sigma} \sum_{j=1}^{p} |\beta_j| \right\} \tag{3.2}$$

for the posterior distribution of $\boldsymbol{\beta}$. From the posterior distribution one can compute Bayes estimators of $\boldsymbol{\beta}$ such as the mean, median or mode and quantify variability using probability intervals. The mode of the posterior distribution of $\boldsymbol{\beta}$ is the value of $\boldsymbol{\beta}$ that maximises $p(\boldsymbol{\beta}|\,\sigma^2, \boldsymbol{y})$. Using equation (3.2),

$$\boldsymbol{\beta}_{mode} = \arg\max_{\boldsymbol{\beta}} \left\{ \exp\left\{ -\frac{(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})}{2\,\sigma^2} - \frac{\gamma}{\sigma}\sum_{j=1}^{p}|\beta_j| \right\} \right\}.$$

With a desire to show that $\boldsymbol{\beta}_{mode}$ is equivalent to $\hat{\boldsymbol{\beta}}_L$, we minimise the negative natural log of the expression. Moreover, when σ^2 is considered to be a constant, we can multiply through the expression by a factor of $2\,\sigma^2$ without changing the solution for $\boldsymbol{\beta}_{mode}$. These steps give

$$\boldsymbol{\beta}_{mode} = \arg\min_{\boldsymbol{\beta}} \left\{ (\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}) + 2\sigma\gamma\sum_{j=1}^{p}|\beta_j| \right\},$$

from which we see that $\boldsymbol{\beta}_{mode}$ is equivalent to $\hat{\boldsymbol{\beta}}_L$ when λ is set equal to $2\sigma\gamma$ (see equation (3.1)).

3.3.2 The Laplace distribution

In order to write the lasso model in a hierarchical structure some convenient distribution theory is employed. The Laplace distribution can be expressed as a scale mixture of normal distributions with independent exponentially distributed variances (Kyung et al., 2010).

Basic Identity. *Normal mixture of the Laplace distribution*

$$\exp\left\{-\frac{\lambda}{\sigma}\,|\beta_j|\right\} \propto \int_0^\infty \underbrace{\left(\frac{1}{2\pi\sigma^2\tau_j^2}\right)^{\frac{1}{2}} \exp\left\{-\frac{\beta_j^2}{2\sigma^2\tau_j^2}\right\}}_{\beta_j \sim N\,(0,\,\sigma^2\tau_j^2)} \underbrace{\left(\frac{\lambda^2}{2}\right)\exp\left\{-\frac{\lambda^2}{2}\tau_j^2\right\}}_{\tau_j^2 \sim Exp\left(\frac{\lambda^2}{2}\right)} d\tau_j^2$$

The Basic Identity introduces p latent parameters $(\tau_1^2, \ldots, \tau_p^2)$ to the model, where the distribution of each β_j, given σ^2 and τ_j^2, is normal. This is desirable, as it yields a closed form expression for the posterior distribution of $\boldsymbol{\beta}$, given all other parameters and the data. This permits computation with the Gibbs sampler.

3.3.3 Hierarchical lasso model

The hierarchical model, as described so far, can be written as:

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

$$\beta_j|\ \sigma^2,\ \tau_j^2 \ \overset{ind}{\sim}\ N(0, \sigma^2\, \tau_j^2), \quad j = 1, \ldots p$$

$$\tau_j^2 \ |\ \lambda^2 \ \overset{ind}{\sim}\ Gamma\left(1,\ \frac{\lambda^2}{2}\right) = Exp\left(\frac{\lambda^2}{2}\right) \quad \tau_j^2 > 0, \quad j = 1, \ldots, p.$$

An inverse-gamma hyperprior can be assumed for σ^2, $\sigma^2 \sim Inv - Gamma(a, b)$. The chosen form of hyperprior maintains conjugacy (Kyung et al., 2010), resulting in a closed expression for the full conditional distribution.

Kyung et al. (2010) report two methods for dealing with the tuning parameter, λ^2, in a Bayesian setting. The most straightforward method is to include λ^2 in the Gibbs sampler. To make this possible, λ^2 is assigned a gamma prior, $\lambda^2 \sim Gamma(\delta, r)$, which maintains conjugacy in the posterior distribution.

Gibbs sampler for the lasso

With all parameters described in the hierarchical model, we have an expression for the joint posterior distribution,

$$p(\boldsymbol{\beta}, \tau_1^2, \ldots, \tau_p^2, \sigma^2, \lambda^2 \mid \boldsymbol{y}) \propto p(\boldsymbol{y}|\boldsymbol{\beta}, \sigma^2) \prod_{j=1}^{p} p(\beta_j | \sigma^2, \tau_j^2) \prod_{j=1}^{p} p(\tau_j^2|\lambda^2) p(\sigma^2 | a, b) p(\lambda^2|\delta, r).$$

The conditional distributions of each parameter, conditional on all other parameters and the data, can be derived and expressed with closed forms from the following distributions (see Kyung et al. (2010, page 400) for the full form of distributions).

- $\boldsymbol{\beta}|\tau_1^2, \ldots, \tau_p^2, \sigma^2, \lambda^2, \boldsymbol{y} \sim$ Multivariate Normal with dimension p
- $\frac{1}{\tau_j^2} |\boldsymbol{\beta}, \sigma^2, \lambda^2, \boldsymbol{y} \overset{ind}{\sim}$ Inverse-Gaussian for $j = 1, \ldots, p$
- $\sigma^2 |\boldsymbol{\beta}, \tau_1^2, \ldots, \tau_p^2, \lambda^2, \boldsymbol{y} \sim$ Inverse-Gamma
- $\lambda^2|\boldsymbol{\beta}, \tau_1^2, \ldots, \tau_p^2, \sigma^2, \boldsymbol{y} \sim$ Gamma

These full conditional distributions allow for computation using the Gibbs sampler as outlined in Section 3.1.2. From the samples of $\boldsymbol{\beta}$, one can obtain estimates of the full probability models of each β_j.

3.4 Discussion

The lasso and its variations are popular approaches for variable selection and estimation. These methods, however, require additional work for computation of standard errors. Bootstrapping is a popular approach for computing standard errors, yet Kyung et al. (2010) have demonstrated that bootstrap errors are not consistent when the true value of the coefficient is zero.

Bayesian inference of the coefficients is based on the estimated posterior distributions. As shown in Section 3.3.1 the point estimate of the usual frequentist lasso is equivalent to the posterior mode of the Bayesian lasso. A key advantage to the Bayesian lasso is that once the Gibbs sampler has been implemented the posterior can be summarised in a variety of ways. It is typical to use the posterior mean as an estimate, but the posterior mode could be used just as easily (Kyung et al., 2010). Additionally, with use of the estimated posterior distributions, credible intervals may be used to assess the sparsity of the model's coefficients.

As a final point, the Bayesian Gibbs samplers proposed here have been tested in a variety of settings and have been shown to either match or outperform their frequentist lasso counterparts when comparing predictive performance (Kyung et al., 2010).

Chapter 4

Group-sparse multi-task regression and feature selection (G-SMuRFS)

4.1 Motivation

Wang et al. (2012) suggest that there are three main challenges faced when attempting to identify quantitative trait loci. i) The data are high dimensional. The human genome consists of thousands of SNPs, and among the many SNPs included in the model only a small fraction are expected to be related to the imaging phenotypes. Accordingly, it is desired to select a subset of SNPs that are relevant to imaging phenotypes. ii) SNPs are connected to QTs through various pathways; multiple SNPs on one gene often jointly carry out genetic functionalities. Therefore, it is desirable to develop a model to exploit the group structure of SNPs. iii) Functionality of the brain involves more than one structure; regarding each QT as a separate outcome ignores the relationship between these structures, which can result in loss of valuable information.

In order to overcome these challenges, Wang et al. (2012) develop *Group-Sparse Multi-task Regression and Feature Selection* (G-SMuRFS) to perform simultaneous estimation and SNP selection across all phenotypes.

4.2 Wang et al. (2012) estimator

The genetic data from ADNI participants takes the form $\boldsymbol{x}_\ell = (x_{\ell 1}, \ldots, x_{\ell d})^T$, for $\ell = 1, \ldots, n$, where n is the number of participants and d is the number of SNPs. Each $x_{\ell i} \in \{0, 1, 2\}$ is the number of minor alleles for the ℓ^{th} subject on the i^{th} SNP. The d SNPs can be partitioned into K genes; π_k, for $k = 1, 2, \ldots, K$. The selected imaging phenotype data is denoted by $\boldsymbol{y}_\ell = (y_{\ell 1}, \ldots, y_{\ell c})^T$, for $\ell = 1, \ldots, n$, where c is the number of imaging phenotypes. Equation (4.1) displays the proposed estimator, where $\boldsymbol{W}$ is a $d \times c$ matrix of regression coefficients.

$$\hat{\boldsymbol{W}} = \arg\min_{\boldsymbol{W}} \left\{ \sum_{\ell=1}^{n} ||\boldsymbol{W}^T \boldsymbol{x}_\ell - \boldsymbol{y}_\ell||_2^2 + \gamma_1 \sum_{k=1}^{K} \sqrt{\sum_{i \in \pi_k} \sum_{j=1}^{c} w_{ij}^2} + \gamma_2 \sum_{i=1}^{d} \sqrt{\sum_{j=1}^{c} w_{ij}^2} \right\}. \quad (4.1)$$

The model consists of three major components. The first component is residual sum of squares. As a result, element w_{ij} of $\boldsymbol{W}$ measures the relative importance of the i^{th} SNP to the j^{th} phenotype. The second component, inspired by group lasso (Yuan & Lin, 2006), introduces a new form of regularisation ($G_{2,1} - norm$) to address group-wise association among SNPs. Coefficients within a group, across all QTs, are penalised together via $\ell_2 - norm$ while $\ell_1 - norm$ is used to sum up group-wise penalties to enforce sparsity between groups. $G_{2,1} - norm$ regularisation differs from group lasso, as it penalises regression coefficients for

a group of SNPs across all responses jointly. As an important group may contain irrelevant individual SNPs, or a less important group may contain individually significant SNPs, the last component is an additional penalty term added for individual structured sparsity. This second penalty term enforces $\ell_{2,1} - norm$ regularisation on individual SNPs. Given $\boldsymbol{X} = (\boldsymbol{x_1}, \ldots, \boldsymbol{x_n})$ and $\boldsymbol{Y} = (\boldsymbol{y_1}, \ldots, \boldsymbol{y_n})$, Equation (4.1) can be rewritten more concisely in matrix form as:

$$\hat{\boldsymbol{W}} = \arg\min_{\boldsymbol{W}} \left\{ ||\boldsymbol{W}^T\boldsymbol{X} - \boldsymbol{Y}||_F^2 + \gamma_1||\boldsymbol{W}||_{G_{2,1}} + \gamma_2||\boldsymbol{W}||_{2,1} \right\}. \tag{4.2}$$

Computation of $\hat{\boldsymbol{W}}$ is based on a simple iterative algorithm that converges to the global optimum. Tuning parameters, γ_1 and γ_2, are chosen by standard 5-fold cross-validation in the range of $(10^{-5}, 10^{-4}, \ldots, 10^4, 10^5)$.

4.3 G-SMuRFS results and limitations

Wang et al. (2012) evaluate their method with an application to data from ADNI. A total of 1224 SNPs are examined, with subsets of SNPs, ranging in size, used to predict the responses of the MRI imaging phenotypes from selected ROIs. Their method is compared against multivariate linear regression, ridge regression, and multi-task feature learning (Evgeniou & Pontil, 2007). When comparing predictive performance, Wang et al. (2012) find the G-SMuRFS method consistently outperforms the three competing methods.

Despite these promising results, the proposed G-SMuRFS method only provides point estimates of regression coefficients and lacks an approach for computing standard errors. By noting the connection between penalised regression methods and Bayesian models (Kyung

et al., 2010; Park & Casella, 2008), we develop an equivalent hierarchical Bayesian model in the next chapter. This eventually allows for inference based on the posterior distributions.

Chapter 5

Development of Bayesian group sparse multi-task regression

In this chapter, we systematically develop the Bayesian formulation of G-SMuRFS. Much of the notation used throughout this chapter is set in Section 5.1. Subsequent notation is defined in each relevant section. We introduce a simple Bayesian model in Section 5.2, and then move to a more convenient, yet equivalent, hierarchical model in Section 5.3. Step by step derivations of full conditional distributions are shown in Section 5.4, and computation is discussed in Section 5.5.

5.1 Notation

- Let m_k be the number of SNPs in the k^{th} gene, π_k, for $k = 1, \ldots, K$.
- $\boldsymbol{W}^{(k)}$ is a matrix consisting of all the rows of $\boldsymbol{W}$ that correspond to the SNPs included

in π_k. The rows that correspond to SNPs that are not included in π_k are removed from $\boldsymbol{W}^{(k)}$ and the remaining rows are kept in the same order. This makes $\boldsymbol{W}^{(k)}$ a $m_k \times c$ matrix.

- $\boldsymbol{W}^{(-k)}$ is a matrix consisting of all the rows of $\boldsymbol{W}$ that correspond to the SNPs not included in π_k. The rows that correspond to SNPs that are included in π_k are removed from $\boldsymbol{W}$ and the remaining rows are kept in the same order. This makes $\boldsymbol{W}^{(-k)}$ a $(d - m_k) \times c$ matrix.

- $\boldsymbol{x}_\ell^{(k)}$ is a vector consisting of all entries of $\boldsymbol{x}_\ell$ (SNP data from the ℓ^{th} subject) that correspond to SNPs included in π_k. The entries of $\boldsymbol{x}_\ell$ that correspond to SNPs that are not included in π_k are removed from $\boldsymbol{x}_\ell^{(k)}$ and the remaining entries are kept in the same order. This makes $\boldsymbol{x}_\ell^{(k)}$ a column vector of length m_k.

- $\boldsymbol{x}_\ell^{(-k)}$ is a vector consisting of all entries of $\boldsymbol{x}_\ell$ that correspond to SNPs not included in π_k. The other entries of $\boldsymbol{x}_\ell$ are removed from $\boldsymbol{x}_\ell^{(k)}$ and the remaining entries are kept in the same order. This makes $\boldsymbol{x}_\ell^{(-k)}$ a column vector of length $(d - m_k)$.

- Let $\boldsymbol{\zeta}^2$ denote the vector $(\zeta_1^2, \ldots, \zeta_{\mathrm{d}}^2)^T$, where ζ_{i}^2 is a parameter associated with the i^{th} SNP.

- Let $\boldsymbol{\tau}^2$ denote the vector $(\tau_1^2, \ldots, \tau_{\mathrm{K}}^2)^T$, where τ_k^2 is a parameter associated with the k^{th} gene.

Definition 3 (Vectorisation of a matrix)**.** ˙

The vectorisation of a matrix, $\boldsymbol{W}$, is a linear transformation of $\boldsymbol{W}$ to a column vector. The columns of $\boldsymbol{W}$ are stacked one under the other to form a single column.

Eg. If $\boldsymbol{W}$ is a $d \times c$ matrix, then vec$(\boldsymbol{W}^T) = [w_{1,1}, \ldots, w_{1,c}, w_{2,1}, \ldots, w_{2,c}, \ldots, w_{d,1}, \ldots, w_{d,c}]^T$.

5.2 Simple Bayesian formulation

Recall the G-SMuRFS estimator,

$$\hat{\boldsymbol{W}} = \arg\min_{\boldsymbol{W}} \left\{ \sum_{\ell=1}^{n} ||\boldsymbol{W}^T \boldsymbol{x}_\ell - \boldsymbol{y}_\ell||_2^2 + \gamma_1 \sum_{k=1}^{K} \sqrt{\sum_{i \in \pi_k} \sum_{j=1}^{c} w_{ij}^2} + \gamma_2 \sum_{i=1}^{d} \sqrt{\sum_{j=1}^{c} w_{ij}^2} \right\}. \quad (5.1)$$

Here we develop a Bayesian model and show that the posterior mode of $\boldsymbol{W}$ is equivalent to the estimator (5.1).

5.2.1 Model A

The quantitative imaging traits, conditional on $\boldsymbol{W}$ and σ^2, are independently distributed as multivariate normal,

$$\boldsymbol{y}_\ell | \boldsymbol{W}, \sigma^2 \overset{ind}{\sim} MVN_c(\boldsymbol{W}^T \boldsymbol{x}_\ell \, , \, \sigma^2 I_c), \quad \ell = 1, \ldots, n. \quad (5.2)$$

We assign conditionally independent priors to $\boldsymbol{W}^{(k)}$,

$$\boldsymbol{W}^{(k)} | \sigma^2, \lambda_1, \lambda_2 \overset{ind}{\sim} p(\boldsymbol{W}^{(k)} | \sigma^2, \lambda_1, \lambda_2), \quad k = 1, \ldots, K, \quad (5.3)$$

with the prior distribution for each $\boldsymbol{W}^{(k)}$ having a density function given by,

$$p(\boldsymbol{W}^{(k)} | \sigma^2, \lambda_1, \lambda_2) \propto \exp\left\{ -\frac{\lambda_1}{\sigma} \sqrt{\sum_{i \in \pi_k} \sum_{j=1}^{c} w_{ij}^2} \right\} \prod_{i \in \pi_k} \exp\left\{ \frac{\lambda_2}{\sigma} \sqrt{\sum_{j=1}^{c} w_{ij}^2} \right\}. \quad (5.4)$$

Proposition 1. *Given λ_1, λ_2, σ^2, the posterior mode of $\boldsymbol{W}$ associated with **Model A** ((5.2) - (5.4)) is exactly $\hat{\boldsymbol{W}}$ from (5.1).*

Proof.

$$
\begin{aligned}
&p(\boldsymbol{W}|\, \boldsymbol{Y}, \sigma^2, \lambda_1, \lambda_2) \\
&\quad \propto p(\boldsymbol{Y}|\, \boldsymbol{W}, \sigma^2) \prod_{k=1}^{K} p(\boldsymbol{W}^{(k)}|\sigma^2, \lambda_1, \lambda_2) \\
&\quad \propto \exp\left\{ -\frac{1}{2\sigma^2} \sum_{\ell=1}^{n} (\boldsymbol{y}_\ell - \boldsymbol{W}^T \boldsymbol{x}_\ell)^T (\boldsymbol{y}_\ell - \boldsymbol{W}^T \boldsymbol{x}_\ell) - \frac{\lambda_1}{\sigma} \sum_{k=1}^{K} \sqrt{\sum_{i \in \pi_k} \sum_{j=1}^{c} w_{ij}^2} - \frac{\lambda_2}{\sigma} \sum_{k=1}^{K} \sum_{i \in \pi_k} \sqrt{\sum_{j=1}^{c} w_{ij}^2} \right\}.
\end{aligned}
$$

The mode of the posterior distribution is equivalent to the $\boldsymbol{W}$ matrix that maximises the above expression, and so it follows that

$$
\boldsymbol{W}_{mode} = \arg\max_{\boldsymbol{W}} \left\{ \exp\left\{ -\frac{1}{2\sigma^2} \sum_{\ell=1}^{n} ||\boldsymbol{W}^T \boldsymbol{x}_\ell - \boldsymbol{y}_\ell||_2^2 - \frac{\lambda_1}{\sigma} \sum_{k=1}^{K} \sqrt{\sum_{i \in \pi_k} \sum_{j=1}^{c} w_{ij}^2} - \frac{\lambda_2}{\sigma} \sum_{i=1}^{d} \sqrt{\sum_{j=1}^{c} w_{ij}^2} \right\} \right\}.
$$

To show that $\boldsymbol{W}_{mode}$ is equivalent to $\hat{\boldsymbol{W}}$ in (5.1), we minimise the negative natural log of the function. This, coupled with some minimal rearranging, gives

$$
\boldsymbol{W}_{mode} = \arg\min_{\boldsymbol{W}} \left\{ \sum_{\ell=1}^{n} ||\boldsymbol{W}^T \boldsymbol{x}_\ell - \boldsymbol{y}_\ell||_2^2 + \, 2\lambda_1 \sigma \sum_{k=1}^{K} \sqrt{\sum_{i \in \pi_k} \sum_{j=1}^{c} w_{ij}^2} + \, 2\lambda_2 \sigma \sum_{i=1}^{d} \sqrt{\sum_{j=1}^{c} w_{ij}^2} \right\},
$$

which shows the mode of the posterior distribution is the same as the proposed $\hat{\boldsymbol{W}}$, where $\gamma_1 = 2\lambda_1\sigma$ and $\gamma_2 = 2\lambda_2\sigma$.

□

5.3 Bayesian hierarchical model

Since it is not convenient to work with the prior $p(\boldsymbol{W} \mid \sigma^2, \lambda_1, \lambda_2)$ from *Model A*, we follow the ideas from Kyung et al. (2010), as outlined in Section 3.3, and develop an equivalent lasso hierarchy.

Let $k : \{1, \ldots, d\} \to \{1, \ldots, K\}$, where $k(i)$ is the gene associated with the i^{th} SNP.

5.3.1 Model B

- $\boldsymbol{y}_\ell \mid \boldsymbol{W},\ \sigma^2 \overset{ind}{\sim} MVN_c\left(\boldsymbol{W}^T \boldsymbol{x}_\ell,\ \sigma^2 I_c\right),\quad \ell = 1, \ldots, n$
- $w_{ij} \mid \sigma^2,\ \boldsymbol{\tau}^2,\ \boldsymbol{\zeta}^2 \overset{ind}{\sim} N\left(0,\ \sigma^2(\frac{1}{\tau^2_{k(i)}} + \frac{1}{\zeta_i^2})^{-1}\right),\quad i = 1, \ldots, d,\ j = 1, \ldots, c$
- $\tau_k^2 \mid \lambda_1 \overset{ind}{\sim} Gamma\left(\frac{m_k c+1}{2},\ \frac{\lambda_1^2}{2}\right),\quad k = 1, \ldots, K$
- $\zeta_i^2 \mid \lambda_2 \overset{ind}{\sim} Gamma\left(\frac{c+1}{2},\ \frac{\lambda_2^2}{2}\right),\quad i = 1, \ldots, d$
- $\boldsymbol{\tau}$ and $\boldsymbol{\zeta}$ are assumed independent
- $\sigma^2 \sim Inv - Gamma(a_\sigma, b_\sigma)$

Proposition 2. *Given σ^2, λ_1, λ_2, **Model A** is equivalent to **Model B**.*

Proof.

Assume *Model A*.

The prior from *Model A*, $p(\boldsymbol{W}|\sigma^2, \lambda_1, \lambda_2)$, is split into its K components, $\prod_{k=1}^K p(\boldsymbol{W}^{(k)}|\sigma^2, \lambda_1, \lambda_2)$,

where

$$p(\boldsymbol{W}^{(k)}|\sigma^2,\lambda_1,\lambda_2) \propto \exp\left\{-\frac{\lambda_1}{\sigma}\sqrt{\sum_{i\in\pi_k}\sum_{j=1}^{c} w_{ij}^2}\right\}\prod_{i\in\pi_k}\exp\left\{\frac{\lambda_2}{\sigma}\sqrt{\sum_{j=1}^{c} w_{ij}^2}\right\}.$$

Let $||\boldsymbol{W}^{(k)}||_2 = \sqrt{\sum_{i\in\pi_k}\sum_{j=1}^{c} w_{ij}^2}$. The K latent parameters, $(\tau_1^2,\ldots,\tau_K^2)$, are introduced to the model using the identity presented in Kyung et al. (2010, page 400).

$$\exp\left\{-\frac{\lambda_1}{\sigma}\,||\boldsymbol{W}^{(k)}||_2\right\} \propto \int_0^\infty \underbrace{\left(\frac{1}{2\pi\sigma^2\tau_k^2}\right)^{\frac{m_k c}{2}}\exp\left\{-\frac{||\boldsymbol{W}^{(k)}||_2^2}{2\sigma^2\tau_k^2}\right\}}_{\mathrm{vec}(\boldsymbol{W}^{(k)T})\sim MVN_{m_k c}\left(\boldsymbol{0},\,\sigma^2\tau_k^2 I_{m_k c}\right)} \underbrace{\frac{\left(\frac{\lambda_1^2}{2}\right)^{\left(\frac{m_k c+1}{2}\right)}}{\Gamma\left(\frac{m_k c+1}{2}\right)}(\tau_k^2)^{\left(\frac{m_k c+1}{2}\right)-1}\exp\left\{-\left(\frac{\lambda_1^2}{2}\right)\tau_k^2\right\}}_{\tau_k^2\sim Gamma\left(\left(\frac{m_k c+1}{2}\right),\left(\frac{\lambda_1^2}{2}\right)\right)} d\tau_k^2. \tag{5.5}$$

Let $\boldsymbol{w}^i$ be the i^{th} row of $\boldsymbol{W}$; it follows that $||\boldsymbol{w}^i||_2 = \sqrt{\sum_{j=1}^{c} w_{ij}^2}$.
Another d latent parameters $(\zeta_1^2,\ldots,\zeta_d^2)$ are introduced using the same identity (Kyung et al., 2010, page 400).

$$\exp\left\{-\frac{\lambda_2}{\sigma}||\boldsymbol{w}^i||_2\right\} \propto \int_0^\infty \underbrace{\left(\frac{1}{2\pi\sigma^2\zeta_i^2}\right)^{\frac{c}{2}}\exp\left\{-\frac{||\boldsymbol{w}^i||_2^2}{2\sigma^2\zeta_i^2}\right\}}_{\boldsymbol{w}^i\sim MVN_c\left(0,\,\sigma^2\zeta_i^2 I_c\right)} \underbrace{\frac{\left(\frac{\lambda_2^2}{2}\right)^{\left(\frac{c+1}{2}\right)}}{\Gamma\left(\frac{c+1}{2}\right)}(\zeta_i^2)^{\left(\frac{c+1}{2}\right)-1}\exp\left\{-\left(\frac{\lambda_2^2}{2}\right)\zeta_i^2\right\}}_{\zeta_i^2\sim Gamma\left(\left(\frac{c+1}{2}\right),\left(\frac{\lambda_2^2}{2}\right)\right)} d\zeta_i^2. \tag{5.6}$$

By combining (5.5) and (5.6), the distribution of $\boldsymbol{W}$ conditional on $\tau_1^2,\ldots,\tau_K^2,\zeta_1^2,\ldots,\zeta_d^2$,

and σ^2 can be expressed as

$$\begin{aligned}
\boldsymbol{W} \mid \sigma^2, \tau_1^2, \ldots, \tau_K^2, \zeta_1^2, \ldots, \zeta_d^2 \propto & \prod_{k=1}^{K} \left[\exp\left\{ -\frac{||\boldsymbol{W}^{(k)}||_2^2}{2\sigma^2\tau_k^2} \right\} \prod_{i \in \pi_k} \exp\left\{ -\frac{||\boldsymbol{w}^i||_2^2}{2\sigma^2\zeta_i^2} \right\} \right] \\
&= \exp\left\{ \sum_{k=1}^{K} \frac{\sum_{i \in \pi_k} \sum_{j=1}^{c} w_{ij}^2}{2\sigma^2\tau_k^2} - \sum_{k=1}^{K} \sum_{i \in \pi_k} \frac{\sum_{j=1}^{c} w_{ij}^2}{2\sigma^2\zeta_i^2} \right\} \\
&= \exp\left\{ -\sum_{\mathrm{k}=1}^{\mathrm{K}} \sum_{\mathrm{i} \in \pi_{\mathrm{k}}} \sum_{j=1}^{c} w_{ij}^2 \cdot \left(\frac{1}{2\sigma^2\tau_k^2} + \frac{1}{2\sigma^2\zeta_i^2} \right) \right\} \\
&= \exp\left\{ -\sum_{i=1}^{d} \sum_{j=1}^{c} w_{ij}^2 \cdot \left(\frac{1}{2\sigma^2\tau_{k(i)}^2} + \frac{1}{2\sigma^2\zeta_i^2} \right) \right\} \\
&= \prod_{i=1}^{d} \prod_{j=1}^{c} \exp\left\{ -\frac{w_{ij}^2}{2\sigma^2 \left(\frac{1}{\tau_{k(i)}^2} + \frac{1}{\zeta_i^2} \right)^{-1}} \right\}.
\end{aligned}$$

This shows that the conditional distribution of w_{ij} corresponds to that of *Model B*, namely,

$$w_{ij} \mid \sigma^2, \underbrace{\tau_1^2, \ldots, \tau_K^2}_{Gene\ specific}, \underbrace{\zeta_1^2, \ldots, \zeta_d^2}_{SNP\ specific} \overset{ind}{\sim} N\left(0,\ \sigma^2 \left(\frac{1}{\tau_{k(i)}^2} + \frac{1}{\zeta_i^2} \right)^{-1} \right), \quad i = 1, \ldots, d,\ j = 1, \ldots, c.$$

The conditional distributions of τ_{k}^2 and ζ_{i}^2 are specified in (5.5) and (5.6) from *Model A*,

$$\tau_k^2 \mid \lambda_1 \overset{ind}{\sim} Gamma\left(\left(\frac{m_k c + 1}{2} \right), \left(\frac{\lambda_1^2}{2} \right) \right), \quad k = 1, \ldots, K$$

and

$$\zeta_i^2 \mid \lambda_2 \overset{ind}{\sim} Gamma\left(\left(\frac{c+1}{2} \right), \left(\frac{\lambda_2^2}{2} \right) \right), \quad i = 1, \ldots, d.$$

This shows that *Model A* is equivalent to *Model B* given σ^2, λ_1^2, and λ_2^2.

□

5.4 Full conditional distributions

The proposed hierarchical model results in full conditional distributions with closed form expressions. The full conditional distributions are derived in this section. We first introduce an additional theorem and definition used for derivations.

Definition 4 (The Kronecker product)**.**

Given that $\boldsymbol{A}$ *is a* $k \times m$ *matrix and* $\boldsymbol{B}$ *is a* $m \times n$ *matrix, the Kronecker product is defined as:*

$$\boldsymbol{A} \otimes \boldsymbol{B} = \begin{pmatrix} a_{1,1}\boldsymbol{B} & \dots & a_{1,m}\boldsymbol{B} \\ \vdots & & \vdots \\ a_{k,1}\boldsymbol{B} & \dots & a_{k,m}\boldsymbol{B} \end{pmatrix}$$

Theorem 2.

$vec(AB) = (B^T \otimes I_k)vec(A)$, where k is the number of rows in A and $\otimes$ denotes the Kronecker product.

5.4.1 Derivations

The joint posterior distribution

$$p(\boldsymbol{W}, \tau_1^2, \cdots, \tau_{\mathrm{K}}^2, \zeta_1^2, \cdots, \zeta_{\mathrm{d}}^2, \sigma^2 \,|\boldsymbol{Y})$$

$$\propto p(\boldsymbol{Y}|\boldsymbol{W}, \sigma^2) \cdot p(\boldsymbol{W}|\, \sigma^2, \boldsymbol{\tau}^2, \boldsymbol{\zeta}^2) p(\boldsymbol{\tau}^2 \mid \lambda_1^2) \cdot p(\boldsymbol{\zeta}^2 \mid \lambda_2^2) \cdot p(\sigma^2 \,|a_\sigma, b_\sigma)$$

$$\begin{aligned}
&= \prod_{\ell=1}^{\mathrm{n}} MVN_c\left(\boldsymbol{y}_\ell \middle| \boldsymbol{W}^T \boldsymbol{x}_\ell, \sigma^2\, I_c\right) \prod_{\mathrm{k}=1}^{\mathrm{K}} \prod_{\mathrm{i} \in \pi_{\mathrm{k}}} \prod_{\mathrm{j}=1}^{\mathrm{c}} N\left(w_{ij} \middle| 0, \sigma^2 \left(\frac{1}{\tau_k^2} \,+\, \frac{1}{\zeta_i^2}\right)^{-1}\right) \\
&\quad \prod_{\mathrm{k}=1}^{\mathrm{K}} Gamma\left(\tau_k^2 \middle| \left(\frac{m_k c + 1}{2}\right), \left(\frac{\lambda_1^2}{2}\right)\right) \prod_{\mathrm{i}=1}^{\mathrm{d}} Gamma\left(\zeta_{\mathrm{i}}^2 \,\middle|\, \left(\frac{c+1}{2}\right), \left(\frac{\lambda_2^2}{2}\right)\right) \\
&\quad Inv - Gamma(\sigma^2 \,|a_\sigma, b_\sigma)
\end{aligned}$$

$$\begin{aligned}
&\propto |\, \sigma^2\, I_c|^{-\frac{n}{2}} \exp\left\{-\frac{1}{2\,\sigma^2} \sum_{\ell=1}^{n} (\boldsymbol{y}_\ell - \boldsymbol{W}^T \boldsymbol{x}_\ell)^T (\boldsymbol{y}_\ell - \boldsymbol{W}^T \boldsymbol{x}_\ell)\right\} \\
&\quad \prod_{\mathrm{k}=1}^{\mathrm{K}} \left[\prod_{i \in \pi_k} \left[\left(\sigma^2 \left(\frac{1}{\tau_k^2} \,+\, \frac{1}{\zeta_i^2}\right)^{-1}\right)^{-\frac{c}{2}}\right] \exp\left\{-\sum_{\mathrm{i} \in \pi_{\mathrm{k}}} \left(\frac{\sum_{\mathrm{j}=1}^{\mathrm{c}} w_{ij}^2}{2\,\sigma^2 \left(\frac{1}{\tau_k^2} \,+\, \frac{1}{\zeta_i^2}\right)^{-1}}\right)\right\}\right] \\
&\quad \prod_{\mathrm{k}=1}^{\mathrm{K}} \left[\frac{\left(\frac{\lambda_1^2}{2}\right)^{\left(\frac{m_k c+1}{2}\right)}}{\Gamma\left(\frac{m_k c+1}{2}\right)} (\tau_k^2)^{\left(\frac{m_k c+1}{2}\right)-1} \exp\left\{-\left(\frac{\lambda_1^2}{2}\right) \tau_k^2\right\}\right] \prod_{\mathrm{i}=1}^{\mathrm{d}} \left[\frac{\left(\frac{\lambda_2^2}{2}\right)^{\left(\frac{c+1}{2}\right)}}{\Gamma\left(\frac{c+1}{2}\right)} (\zeta_{\mathrm{i}}^2)^{\left(\frac{c+1}{2}\right)-1} \exp\left\{-\left(\frac{\lambda_2^2}{2}\right) \zeta_{\mathrm{i}}^2\right\}\right] \\
&\quad \frac{b_\sigma^{a_\sigma}}{\Gamma(a_\sigma)} (\sigma^2)^{-a_\sigma - 1} \exp\left\{-\frac{b_\sigma}{\sigma^2}\right\}.
\end{aligned}$$

Full conditional distribution of $\boldsymbol{W}^{(k)}$

$$p(\boldsymbol{W}^{(k)}\big|\boldsymbol{Y},\boldsymbol{W}^{(-k)},\boldsymbol{\tau}^2,\boldsymbol{\zeta}^2,\sigma^2,\lambda_1^2,\lambda_2^2) \propto$$
$$\exp\left\{\frac{-1}{2\,\sigma^2}\sum_{\ell=1}^{n}(\boldsymbol{y}_\ell-\boldsymbol{W}^T\boldsymbol{x}_\ell)^T(\boldsymbol{y}_\ell-\boldsymbol{W}^T\boldsymbol{x}_\ell)\right\}\exp\left\{-\sum_{\mathrm{i}\in\pi_\mathrm{k}}\left(\frac{\sum_{\mathrm{j}=1}^{\mathrm{c}}w_{ij}^2}{2\,\sigma^2\left(\frac{1}{\tau_k^2}\,+\,\frac{1}{\zeta_i^2}\right)^{-1}}\right)\right\}. \quad (5.7)$$

Split $\boldsymbol{W}$ into $\boldsymbol{W}^{(k)}$ and $\boldsymbol{W}^{(-k)}$ and rewrite the first exponent of (5.7).

$$\exp\left\{\frac{-1}{2\,\sigma^2}\sum_{\ell=1}^{\mathrm{n}}(\boldsymbol{y}_\ell-\boldsymbol{W}^{(k)T}\boldsymbol{x}_\ell^{(k)}-\boldsymbol{W}^{(-k)T}\boldsymbol{x}_\ell^{(-k)})^T(\boldsymbol{y}_\ell-\boldsymbol{W}^{(k)T}\boldsymbol{x}_\ell^{(k)}-\boldsymbol{W}^{(-k)T}\boldsymbol{x}_\ell^{(-k)})\right\} \quad (5.8)$$

Theorem 2 is used to vectorise terms that include either $\boldsymbol{W}^{(k)}$ or $\boldsymbol{W}^{(-k)}$.

1) $\mathrm{vec}(\boldsymbol{W}^{(k)T}\boldsymbol{x}_\ell^{(k)}) = (\boldsymbol{x}_\ell^{(k)T}\otimes I_c)\mathrm{vec}(\boldsymbol{W}^{(k)T})$.

2) $\mathrm{vec}(\boldsymbol{W}^{(-k)T}\boldsymbol{x}_\ell^{(-k)}) = (\boldsymbol{x}_\ell^{(-k)T}\otimes I_c)\mathrm{vec}(\boldsymbol{W}^{(-k)T})$.

These results give an equivalent expression for (5.8).

$$\exp\left\{\frac{-1}{2\,\sigma^2}\sum_{\ell=1}^{\mathrm{n}}\Big(\boldsymbol{y}_\ell-(\boldsymbol{x}_\ell^{(k)T}\otimes I_c)\mathrm{vec}(\boldsymbol{W}^{(k)T})-(\boldsymbol{x}_\ell^{(-k)T}\otimes I_c)\mathrm{vec}(\boldsymbol{W}^{(-k)T})\Big)^T\right.$$
$$\left.\Big(\boldsymbol{y}_\ell-(\boldsymbol{x}_\ell^{(k)T}\otimes I_c)\mathrm{vec}(\boldsymbol{W}^{(k)T})-(\boldsymbol{x}_\ell^{(-k)T}\otimes I_c)\mathrm{vec}(\boldsymbol{W}^{(-k)T})\Big)\right\}$$

Apply the transpose to the first section.

$$\exp\left\{\frac{-1}{2\,\sigma^2}\sum_{\ell=1}^{\mathrm{n}}\Big(\boldsymbol{y}_\ell^T-\mathrm{vec}(\boldsymbol{W}^{(k)T})^T(\boldsymbol{x}_\ell^{(k)T}\otimes I_c)^T-\mathrm{vec}(\boldsymbol{W}^{(-k)T})^T(\boldsymbol{x}_\ell^{(-k)T}\otimes I_c)^T\Big)\right.$$
$$\left.\Big(\boldsymbol{y}_\ell-(\boldsymbol{x}_\ell^{(k)T}\otimes I_c)\mathrm{vec}(\boldsymbol{W}^{(k)T})-(\boldsymbol{x}_\ell^{(-k)T}\otimes I_c)\mathrm{vec}(\boldsymbol{W}^{(-k)T})\Big)\right\}$$

Using $(A \otimes B)^T = (A^T \otimes B^T)$ the above is simplified to

$$\exp\left\{\frac{-1}{2\,\sigma^2}\sum_{\ell=1}^{\mathrm{n}}\Big(\boldsymbol{y}_\ell^T - \mathrm{vec}(\boldsymbol{W}^{(k)T})^T(\boldsymbol{x}_\ell^{(k)}\otimes I_c) - \mathrm{vec}(\boldsymbol{W}^{(-k)T})^T(\boldsymbol{x}_\ell^{(-k)}\otimes I_c)\Big)\right.$$
$$\left.\Big(\boldsymbol{y}_\ell - (\boldsymbol{x}_\ell^{(k)T}\otimes I_c)\mathrm{vec}(\boldsymbol{W}^{(k)T}) - (\boldsymbol{x}_\ell^{(-k)T}\otimes I_c)\mathrm{vec}(\boldsymbol{W}^{(-k)T})\Big)\right\}.$$

It is now possible to expand the expression. Only those terms that include $\boldsymbol{W}^{(k)}$ are kept, as the other terms are considered to be constants that can be factored out to become part of the normalising constant.

$$\exp\left\{\frac{-1}{2\,\sigma^2}\sum_{\ell=1}^{\mathrm{n}}\Big(-\boldsymbol{y}_\ell^T(\boldsymbol{x}_\ell^{(k)T}\otimes I_c)\mathrm{vec}(\boldsymbol{W}^{(k)T}) - \mathrm{vec}(\boldsymbol{W}^{(k)T})^T(\boldsymbol{x}_\ell^{(k)}\otimes I_c)\boldsymbol{y}_\ell\right.$$
$$+\mathrm{vec}(\boldsymbol{W}^{(k)T})^T(\boldsymbol{x}_\ell^{(k)}\otimes I_c)(\boldsymbol{x}_\ell^{(k)T}\otimes I_c)\mathrm{vec}(\boldsymbol{W}^{(k)T})$$
$$+\mathrm{vec}(\boldsymbol{W}^{(k)T})^T(\boldsymbol{x}_\ell^{(k)}\otimes I_c)(\boldsymbol{x}_\ell^{(-k)T}\otimes I_c)\mathrm{vec}(\boldsymbol{W}^{(-k)T})$$
$$\left.+\ \mathrm{vec}(\boldsymbol{W}^{(-k)T})^T(\boldsymbol{x}_\ell^{(-k)}\otimes I_c)(\boldsymbol{x}_\ell^{(k)T}\otimes I_c)\mathrm{vec}(\boldsymbol{W}^{(k)T})\Big)\right\}$$

Since each term is a scalar, the transpose of some products can be taken to combine terms to give

$$\exp\left\{\frac{-1}{2\,\sigma^2}\left[\mathrm{vec}(\boldsymbol{W}^{(k)T})^T\sum_{\ell=1}^{\mathrm{n}}(\boldsymbol{x}_\ell^{(k)}\otimes I_c)(\boldsymbol{x}_\ell^{(k)T}\otimes I_c)\mathrm{vec}(\boldsymbol{W}^{(k)T})\right.\right.$$
$$\left.\left.+2\mathrm{vec}(\boldsymbol{W}^{(k)T})^T\sum_{\ell=1}^{\mathrm{n}}(\boldsymbol{x}_\ell^{(k)}\otimes I_c)(\boldsymbol{x}_\ell^{(-k)T}\otimes I_c)\mathrm{vec}(\boldsymbol{W}^{(-k)T}) - 2\mathrm{vec}(\boldsymbol{W}^{(k)T})^T\sum_{\ell=1}^{\mathrm{n}}(\boldsymbol{x}_\ell^{(k)}\otimes I_c)\boldsymbol{y}_\ell\right]\right\}.$$

Next, consider the second exponent of (5.7),

$$\exp\left\{\frac{-1}{2\,\sigma^2}\sum_{\mathrm{i}\in\pi_\mathrm{k}}\frac{\sum_{\mathrm{j}=1}^{\mathrm{c}} w_{ij}^2}{\left(\frac{1}{\tau_k^2}\;+\;\frac{1}{\zeta_i^2}\right)^{-1}}\right\}.$$

We define a matrix, $\boldsymbol{H}_k$, such that $\boldsymbol{H}_k = \left[diag\left\{\frac{1}{\tau_\mathrm{k}^2}+\frac{1}{\zeta_\mathrm{i}^2}\right\}_{i\in\pi_k}\otimes I_c\right]$.

Notice that,

$$\sum_{\mathrm{i}\in\pi_\mathrm{k}}\frac{\sum_{\mathrm{j}=1}^{\mathrm{c}} w_{ij}^2}{\left(\frac{1}{\tau_k^2}\;+\;\frac{1}{\zeta_i^2}\right)^{-1}} = \mathrm{vec}(\boldsymbol{W}^{(k)T})^T\boldsymbol{H}_k\mathrm{vec}(\boldsymbol{W}^{(k)T}).$$

We rewrite the expression of (5.7), up to its normalising constant, to get,

$$\begin{aligned}
&p(\boldsymbol{W}^{(k)}\big|\boldsymbol{Y},\boldsymbol{W}^{(-k)},\boldsymbol{\tau}^2,\boldsymbol{\zeta}^2,\sigma^2,\lambda_1^2,\lambda_2^2)\propto\\
&\exp\left\{\frac{-1}{2\,\sigma^2}\left[\mathrm{vec}(\boldsymbol{W}^{(k)T})^T\sum_{\ell=1}^{\mathrm{n}}(\boldsymbol{x}_\ell^{(k)}\otimes I_c)(\boldsymbol{x}_\ell^{(k)T}\otimes I_c)\mathrm{vec}(\boldsymbol{W}^{(k)T})\;+\mathrm{vec}(\boldsymbol{W}^{(k)T})^T\boldsymbol{H}_k\mathrm{vec}(\boldsymbol{W}^{(k)T})\right.\right.\\
&\qquad+2\mathrm{vec}(\boldsymbol{W}^{(k)T})^T\sum_{\ell=1}^{\mathrm{n}}(\boldsymbol{x}_\ell^{(k)}\otimes I_c)(\boldsymbol{x}_\ell^{(-k)T}\otimes I_c)\mathrm{vec}(\boldsymbol{W}^{(-k)T})\\
&\qquad\left.\left.-2\mathrm{vec}(\boldsymbol{W}^{(k)T})^T\sum_{\ell=1}^{\mathrm{n}}(\boldsymbol{x}_\ell^{(k)}\otimes I_c)\boldsymbol{y}_\ell\right]\right\}. \quad (5.9)
\end{aligned}$$

After expanding, the exponent of the multivariate normal distribution is of the form,

$$\exp\left\{-\frac{1}{2}\left[\mathrm{vec}(\boldsymbol{W}^{(k)T})^T\boldsymbol{\Sigma}_k^{-1}\mathrm{vec}(\boldsymbol{W}^{(k)T})-2\mathrm{vec}(\boldsymbol{W}^{(k)T})^T\boldsymbol{\Sigma}_k^{-1}\,\boldsymbol{\mu}_\mathrm{k}+constant\;\right]\right\}. \quad (5.10)$$

Expression (5.9) is a quadratic function of $\mathrm{vec}(\boldsymbol{W}^{(k)T})$ in the exponent. Therefore, the full conditional distribution of $\mathrm{vec}(\boldsymbol{W}^{(k)T})$ is proportional to a multivariate normal distribution of dimension m_kc, with parameters $\boldsymbol{\mu}_\mathrm{k}$ and $\boldsymbol{\Sigma}_k$. The next steps involve matching (5.9) to (5.10).

Solving for $\boldsymbol{\Sigma}_k$

To isolate $\boldsymbol{\Sigma}_k$, consider the parts of (5.9) that have the terms $\exp\left\{-\frac{1}{2}\left[\text{vec}(\boldsymbol{W}^{(k)T})^T\boldsymbol{\Sigma}_k^{-1}\text{vec}(\boldsymbol{W}^{(k)T})\right]\right\}$.
Currently we have

$$\exp\left\{\frac{-1}{2\,\sigma^2}\left[\text{vec}(\boldsymbol{W}^{(k)T})^T\sum_{\ell=1}^{\text{n}}(\boldsymbol{x}_\ell^{(k)}\otimes I_c)(\boldsymbol{x}_\ell^{(k)T}\otimes I_c)\text{vec}(\boldsymbol{W}^{(k)T})+\text{vec}(\boldsymbol{W}^{(k)T})^T\boldsymbol{H}_k\text{vec}(\boldsymbol{W}^{(k)T})\right]\right\}.$$

Rearrange.

$$\exp\left\{-\frac{1}{2}\left[\text{vec}(\boldsymbol{W}^{(k)T})^T\left(\frac{1}{\sigma^2}\left(\sum_{\ell=1}^{\text{n}}(\boldsymbol{x}_\ell^{(k)}\otimes I_c)(\boldsymbol{x}_\ell^{(k)T}\otimes I_c)+\boldsymbol{H}_k\right)\right)\text{vec}(\boldsymbol{W}^{(k)T})\right]\right\}$$

We now observe that

$$\begin{aligned}\boldsymbol{\Sigma}_k^{-1}=&\frac{1}{\sigma^2}\left(\sum_{\ell=1}^{\text{n}}(\boldsymbol{x}_\ell^{(k)}\otimes I_c)(\boldsymbol{x}_\ell^{(k)T}\otimes I_c)+\boldsymbol{H}_k\right),\\ \boldsymbol{\Sigma}_k=&\,\sigma^2\left(\sum_{\ell=1}^{\text{n}}(\boldsymbol{x}_\ell^{(k)}\otimes I_c)(\boldsymbol{x}_\ell^{(k)T}\otimes I_c)+\boldsymbol{H}_k\right)^{-1}.\end{aligned}$$

This gives the result $\boldsymbol{\Sigma}_k=\sigma^2\,\boldsymbol{A}_k^{-1}$,
where $\boldsymbol{A}_k=\left(\sum_{\ell=1}^{\text{n}}(\boldsymbol{x}_\ell^{(k)}\otimes I_c)(\boldsymbol{x}_\ell^{(k)T}\otimes I_c)+\left(diag\left\{\frac{1}{\tau_{\text{k}}^2}+\frac{1}{\zeta_{\text{i}}^2}\right\}_{i\in\pi_k}\otimes I_c\right)\right)$.

Solving for $\boldsymbol{\mu}_{\mathrm{k}}$

The exponent of the form $-\frac{1}{2}\left(-2\mathrm{vec}(\boldsymbol{W}^{(k)T})^T\boldsymbol{\Sigma}_k^{-1}\,\boldsymbol{\mu}_{\mathrm{k}}\right)$ from the multivariate normal distribution is considered. We have the expression,

$$
\begin{aligned}
&-\frac{1}{2\,\sigma^2}\left(2\mathrm{vec}(\boldsymbol{W}^{(k)T})^T\sum_{\ell=1}^{\mathrm{n}}(\boldsymbol{x}_\ell^{(k)}\otimes I_c)(\boldsymbol{x}_\ell^{(-k)T}\otimes I_c)\mathrm{vec}(\boldsymbol{W}^{(-k)T})-2\mathrm{vec}(\boldsymbol{W}^{(k)T})^T\sum_{\ell=1}^{\mathrm{n}}(\boldsymbol{x}_\ell^{(k)}\otimes I_c)\boldsymbol{y}_\ell\right)\\
&=\mathrm{vec}(\boldsymbol{W}^{(k)T})^T\left(\frac{1}{\sigma^2}\left(-\sum_{\ell=1}^{\mathrm{n}}(\boldsymbol{x}_\ell^{(k)}\otimes I_c)(\boldsymbol{x}_\ell^{(-k)T}\otimes I_c)\mathrm{vec}(\boldsymbol{W}^{(-k)T})+\sum_{\ell=1}^{\mathrm{n}}(\boldsymbol{x}_\ell^{(k)}\otimes I_c)\boldsymbol{y}_\ell\right)\right).
\end{aligned}
$$

Match up the expressions.

$$
\boldsymbol{\Sigma}_k^{-1}\,\boldsymbol{\mu}_{\mathrm{k}}=\frac{1}{\sigma^2}\left(-\sum_{\ell=1}^{\mathrm{n}}(\boldsymbol{x}_\ell^{(k)}\otimes I_c)(\boldsymbol{x}_\ell^{(-k)T}\otimes I_c)\mathrm{vec}(\boldsymbol{W}^{(-k)T})+\sum_{\ell=1}^{\mathrm{n}}(\boldsymbol{x}_\ell^{(k)}\otimes I_c)\boldsymbol{y}_\ell\right).
$$

Isolate $\boldsymbol{\mu}_{\mathrm{k}}$.

$$
\begin{aligned}
\boldsymbol{\mu}_{\mathrm{k}}=&\boldsymbol{\Sigma}_k\left(\frac{1}{\sigma^2}\left(-\sum_{\ell=1}^{\mathrm{n}}(\boldsymbol{x}_\ell^{(k)}\otimes I_c)(\boldsymbol{x}_\ell^{(-k)T}\otimes I_c)\mathrm{vec}(\boldsymbol{W}^{(-k)T})+\sum_{\ell=1}^{\mathrm{n}}(\boldsymbol{x}_\ell^{(k)}\otimes I_c)\boldsymbol{y}_\ell\right)\right)\\
=&\boldsymbol{A}_k^{-1}\left(-\sum_{\ell=1}^{\mathrm{n}}(\boldsymbol{x}_\ell^{(k)}\otimes I_c)(\boldsymbol{x}_\ell^{(-k)T}\otimes I_c)\mathrm{vec}(\boldsymbol{W}^{(-k)T})+\sum_{\ell=1}^{\mathrm{n}}(\boldsymbol{x}_\ell^{(k)}\otimes I_c)\boldsymbol{y}_\ell\right).
\end{aligned}
$$

Finally, the full conditional distribution of $\boldsymbol{W^{(k)}}$ is expressed as

$$
\mathrm{vec}(\boldsymbol{W}^{(k)T})\big|\boldsymbol{Y},\boldsymbol{W}^{(-k)},\boldsymbol{\tau}^2,\boldsymbol{\zeta}^2,\sigma^2,\lambda_1^2,\lambda_2^2\sim MVN_{m_kc}(\,\boldsymbol{\mu}_{\mathrm{k}},\;\boldsymbol{\Sigma}_k),
$$

where

$$
\boldsymbol{\mu}_{\mathrm{k}}=\boldsymbol{A}_k^{-1}\left(-\sum_{\ell=1}^{\mathrm{n}}(\boldsymbol{x}_\ell^{(k)}\otimes I_c)(\boldsymbol{x}_\ell^{(-k)T}\otimes I_c)\mathrm{vec}(\boldsymbol{W}^{(-k)T})+\sum_{\ell=1}^{\mathrm{n}}(\boldsymbol{x}_\ell^{(k)}\otimes I_c)\boldsymbol{y}_\ell\right),
$$

$$\boldsymbol{A}_k = \left(\sum_{\ell=1}^{\mathrm{n}} (\boldsymbol{x}_\ell^{(k)} \otimes I_c)(\boldsymbol{x}_\ell^{(k)T} \otimes I_c) + \left(diag \left\{ \frac{1}{\tau_{\mathrm{k}}^2} + \frac{1}{\zeta_{\mathrm{i}}^2} \right\}_{i \in \pi_k} \otimes I_c \right) \right), \text{ and } \boldsymbol{\Sigma}_k = \sigma^2 \, \boldsymbol{A}_k^{-1}.$$

Full conditional distribution of σ^2

$$p(\sigma^2 \,|\boldsymbol{Y}, \boldsymbol{W}, \boldsymbol{\tau}^2, \boldsymbol{\zeta}^2, \lambda_1^2, \lambda_2^2)$$

$$\propto |\, \sigma^2 \, I_c|^{-\frac{n}{2}} \exp\left\{ -\frac{1}{2\,\sigma^2} \sum_{\ell=1}^{n} (\boldsymbol{y}_\ell - \boldsymbol{W}^T \boldsymbol{x}_\ell)^T (\boldsymbol{y}_\ell - \boldsymbol{W}^T \boldsymbol{x}_\ell) \right\}$$
$$\prod_{\mathrm{k}=1}^{\mathrm{K}} \left[(\sigma^2)^{-\frac{m_k c}{2}} \prod_{\mathrm{i} \in \pi_{\mathrm{k}}} \left[\left(\frac{1}{\tau_k^2} \, + \, \frac{1}{\zeta_i^2} \right)^{-1} \right]^{-\frac{c}{2}} \exp\left\{ -\frac{1}{2\,\sigma^2} \sum_{\mathrm{i} \in \pi_{\mathrm{k}}} \frac{\sum_{\mathrm{j}=1}^{\mathrm{c}} w_{ij}^2}{\left(\frac{1}{\tau_k^2} \, + \, \frac{1}{\zeta_i^2} \right)^{-1}} \right\} \right] \cdot (\sigma^2)^{-a_\sigma - 1} \exp\left\{ -\frac{b_\sigma}{\sigma^2} \right\}$$

$$= \prod_{\mathrm{k}=1}^{\mathrm{K}} \prod_{\mathrm{i} \in \pi_{\mathrm{k}}} \left[\left(\frac{1}{\tau_k^2} \, + \, \frac{1}{\zeta_i^2} \right)^{-1} \right]^{-\frac{c}{2}} \left(\sigma^2\right)^{-\frac{cn}{2}} \left(\sigma^2\right)^{-\frac{dc}{2}} \left(\sigma^2\right)^{-a_\sigma - 1}$$
$$\exp\left\{ -\frac{1}{2\,\sigma^2} \sum_{\ell=1}^{n} ||\boldsymbol{y}_\ell - \boldsymbol{W}^T \boldsymbol{x}_\ell||_2^2 - \frac{1}{2\,\sigma^2} \sum_{\mathrm{i}=1}^{\mathrm{d}} \frac{\sum_{\mathrm{j}=1}^{\mathrm{c}} w_{ij}^2}{\left(\frac{1}{\tau_{k(i)}^2} \, + \, \frac{1}{\zeta_i^2} \right)^{-1}} - \frac{b_\sigma}{\sigma^2} \right\}.$$

Since $\prod_{\mathrm{k}=1}^{\mathrm{K}} \prod_{\mathrm{i} \in \pi_{\mathrm{k}}} \left[\left(\frac{1}{\tau_k^2} \, + \, \frac{1}{\zeta_i^2} \right)^{-1} \right]^{-\frac{c}{2}}$ does not depend on σ^2, it can be factored out of the expression. This step leaves,

$$p(\sigma^2 \,|\boldsymbol{Y}, \boldsymbol{W}, \boldsymbol{\tau}^2, \boldsymbol{\zeta}^2, \lambda_1^2, \lambda_2^2) \propto$$
$$(\sigma^2)^{-\left(\frac{cn}{2} + \frac{dc}{2} + a_\sigma\right) - 1} \exp\left\{ -\frac{1}{\sigma^2} \left(\frac{1}{2} \sum_{\ell=1}^{n} ||\boldsymbol{y}_\ell - \boldsymbol{W}^T \boldsymbol{x}_\ell||_2^2 + \frac{1}{2} \sum_{\mathrm{i}=1}^{\mathrm{d}} \frac{\sum_{\mathrm{j}=1}^{\mathrm{c}} w_{ij}^2}{\left(\frac{1}{\tau_{k(i)}^2} \, + \, \frac{1}{\zeta_i^2} \right)^{-1}} + b_\sigma \right) \right\},$$

and gives the result

$$\sigma^2 \left| \boldsymbol{Y}, \boldsymbol{W}, \boldsymbol{\tau}^2, \boldsymbol{\zeta}^2, \lambda_1^2, \lambda_2^2 \sim \quad Inv - Gamma\left(a_\sigma^*, b_\sigma^*\right), \right.$$

$$\text{where } \; a_\sigma^* = \left(\frac{cn}{2} + \frac{dc}{2} + a_\sigma\right), \; b_\sigma^* = \left(\frac{1}{2}\sum_{\ell=1}^{n} ||\boldsymbol{y}_\ell - \boldsymbol{W}^T \boldsymbol{x}_\ell||_2^2 + \frac{1}{2}\sum_{\mathrm{i}=1}^{\mathrm{d}} \frac{\sum_{\mathrm{j}=1}^{\mathrm{c}} w_{ij}^2}{\left(\frac{1}{\tau_{k(i)}^2} \; + \; \frac{1}{\zeta_i^2}\right)^{-1}} + b_\sigma\right).$$

Full Conditional of τ_k^2

To derive the full conditional distribution of τ_{k}^2, we look at an earlier version of the conditional distribution of $\boldsymbol{W}^{(k)}$ from (5.5).

$$\exp\left\{-\frac{\lambda_1}{\sigma}\,||\boldsymbol{W}^{(k)}||_2\right\}$$
$$\propto \int_0^\infty \underbrace{\left(\frac{1}{2\pi\sigma^2\tau_k^2}\right)^{\frac{m_k c}{2}} \exp\left\{-\frac{||\boldsymbol{W}^{(k)}||_2^2}{2\sigma^2\tau_k^2}\right\}}_{\mathrm{vec}(\boldsymbol{W}^{(k)T}) \sim MVN_{m_k c}\left(\boldsymbol{0}, \sigma^2\tau_k^2 I_{m_k c}\right)} \underbrace{\frac{\left(\frac{\lambda_1^2}{2}\right)^{\left(\frac{m_k c+1}{2}\right)}}{\Gamma\left(\frac{m_k c+1}{2}\right)} (\tau_k^2)^{\left(\frac{m_k c+1}{2}\right)-1} \exp\left\{-\left(\frac{\lambda_1^2}{2}\right)\tau_k^2\right\}}_{\tau_{\mathrm{k}}^2 \sim Gamma\left(\left(\frac{m_k c+1}{2}\right), \left(\frac{\lambda_1^2}{2}\right)\right)} d\tau_k^2.$$

The parameter τ_{k}^2 does not appear anywhere else in the joint posterior distribution. Therefore we have,

$$p(\tau_{\mathrm{k}}^2 \left| \boldsymbol{Y}, \boldsymbol{W}, \boldsymbol{\tau}^{\boldsymbol{2}}_{\boldsymbol{(-k)}}, \boldsymbol{\zeta}^2, \sigma^2, \lambda_2^2, \lambda_1^2\right.) \propto (\tau_{\mathrm{k}}^2)^{-\frac{m_k c}{2}} \exp\left\{-\frac{||\boldsymbol{W}^{(k)}||_2^2}{2\sigma^2\tau_k^2}\right\} (\tau_{\mathrm{k}}^2)^{\left(\frac{m_k c+1}{2}\right)-1} \exp\left\{-\frac{\lambda_1^2}{2}\,\tau_{\mathrm{k}}^2\right\}$$
$$= (\tau_{\mathrm{k}}^2)^{-\frac{1}{2}} \exp\left\{-\frac{||\boldsymbol{W}^{(k)}||_2^2}{2\sigma^2\tau_k^2} - \frac{\lambda_1^2}{2}\,\tau_{\mathrm{k}}^2\right\}.$$

Let $\nu_k = (\tau_{\mathrm{k}}^2)^{-1}$, and Jacobian $= \left|\frac{d}{d\nu_k}\tau_k^2(\nu_k)\right| = \nu_k^{-2}$ to perform a standard change of variables

and obtain the conditional distribution of $\nu_k = \frac{1}{\tau_k^2}$.

$$p(\nu_k \big| \boldsymbol{Y}, \boldsymbol{W}, \boldsymbol{\tau}^{\boldsymbol{2}}_{\boldsymbol{(-k)}}, \boldsymbol{\zeta}^2, \sigma^2, \lambda_2^2, \lambda_1^2) \propto (\nu_k^{-1})^{-\frac{1}{2}} \nu_k^{-2} \exp\left\{ -\frac{||\boldsymbol{W}^{(k)}||_2^2}{2\sigma^2 \nu_k^{-1}} - \frac{\lambda_1^2}{2} \nu_k^{-1} \right\}$$
$$= \nu_k^{-\frac{3}{2}} \exp\left\{ -\frac{||\boldsymbol{W}^{(k)}||_2^2 \nu_k}{2\,\sigma^2} - \frac{\lambda_1^2}{2\nu_k} \right\}.$$

The *Inverse Gaussian distribution* is parametrised with mean, μ, and shape, κ, and has a probability density function of the form,

$$f(x; \mu, \kappa) = \left(\frac{\kappa}{2\pi x^3} \right)^{\frac{1}{2}} \exp\left\{ \frac{-\kappa(x-\mu)^2}{2\mu^2 x} \right\}.$$

By expanding the probability density function we get the equivalent form,

$$f(x; \mu, \kappa) = \left(\frac{\kappa}{2\pi x^3} \right)^{\frac{1}{2}} \exp\left\{ -\frac{\kappa}{2} \left(\frac{x}{\mu^2} - \frac{2}{\mu} + \frac{1}{x} \right) \right\}.$$

The conditional distribution of ν_k, as seen above, can be manipulated to have the form

$$p(\nu_k \big| \boldsymbol{Y}, \boldsymbol{W}, \boldsymbol{\tau}^{\boldsymbol{2}}_{\boldsymbol{(-k)}}, \boldsymbol{\zeta}^2, \sigma^2, \lambda_2^2, \lambda_1^2) \propto \left(\frac{1}{\nu_k^3} \right)^{\frac{1}{2}} \exp\left\{ -\frac{\lambda_1^2}{2} \left(\frac{\nu_k ||\boldsymbol{W}^{(k)}||_2^2}{\lambda_1^2\,\sigma^2} + \frac{1}{\nu_k} \right) \right\}.$$

The conditional distribution of ν_k follows an Inverse Gaussian distribution.

$$\nu_k = \frac{1}{\tau_k^2} \;\Big|\; \boldsymbol{Y}, \boldsymbol{W}, \boldsymbol{\tau}^{\boldsymbol{2}}_{\boldsymbol{(-k)}}, \boldsymbol{\zeta}^2, \sigma^2, \lambda_1^2, \lambda_2^2 \sim \textit{Inverse-Gaussian}\left(\sqrt{\frac{\lambda_1^2\,\sigma^2}{||\boldsymbol{W}^{(k)}||_2^2}}\,,\;\; \lambda_1^2 \right).$$

Full conditional distribution of ζ_{i}^2

The derivation of the full conditional distribution of ζ_{i}^2 follows the same steps as those of τ_{k}^2. Expression (5.6),

$$\exp\left\{-\frac{\lambda_2}{\sigma}||\boldsymbol{w}^i||_2\right\}$$
$$\propto \int_0^\infty \underbrace{\left(\frac{1}{2\pi\sigma^2\zeta_i^2}\right)^{\frac{c}{2}} \exp\left\{-\frac{||\boldsymbol{w}^i||_2^2}{2\sigma^2\zeta_i^2}\right\}}_{\boldsymbol{w}^i \sim MVN_c\left(0,\,\sigma^2\zeta_i^2 I_c\right)} \underbrace{\frac{\left(\frac{\lambda_2^2}{2}\right)^{\left(\frac{c+1}{2}\right)}}{\Gamma\left(\frac{c+1}{2}\right)}(\zeta_i^2)^{\left(\frac{c+1}{2}\right)-1}\ \exp\left\{-\left(\frac{\lambda_2^2}{2}\right)\zeta_i^2\right\}}_{\zeta_{\mathrm{i}}^2 \sim Gamma\left(\left(\frac{c+1}{2}\right),\left(\frac{\lambda_2^2}{2}\right)\right)} d\zeta_i^2,$$

is used to obtain an initial proportional conditional distribution of ζ_{i}^2,

$$p(\zeta_{\mathrm{i}}^2 \,|\boldsymbol{Y}, \boldsymbol{W}, \boldsymbol{\tau}^2, \boldsymbol{\zeta}^{\boldsymbol{2}}_{\boldsymbol{(-i)}}, \sigma^2, \lambda_1^2, \lambda_2^2) \propto (\zeta_{\mathrm{i}}^2)^{-\frac{c}{2}} \exp\left\{-\frac{||\boldsymbol{w}^i||_2^2}{2\,\sigma^2\,\zeta_{\mathrm{i}}^2}\right\}(\zeta_{\mathrm{i}}^2)^{\left(\frac{c+1}{2}\right)-1} \exp\left\{-\frac{\lambda_2^2}{2}\,\zeta_{\mathrm{i}}^2\right\}$$
$$= (\zeta_{\mathrm{i}}^2)^{-\frac{1}{2}} \exp\left\{-\frac{||\boldsymbol{w}^i||_2^2}{2\,\sigma^2\,\zeta_{\mathrm{i}}^2} - \frac{\lambda_2^2}{2}\,\tau_{\mathrm{k}}^2\right\}.$$

A change of variables is used to find the distribution of $\eta_i = \frac{1}{\zeta_{\mathrm{i}}^2}$,

$$p(\eta_i = \frac{1}{\zeta_{\mathrm{i}}^2}|\boldsymbol{Y}, \boldsymbol{W}, \boldsymbol{\tau}^2, \boldsymbol{\zeta}^{\boldsymbol{2}}_{\boldsymbol{(-i)}}, \sigma^2, \lambda_1^2, \lambda_2^2) \propto \eta_i^{-\frac{3}{2}} \exp\left\{-\frac{||\boldsymbol{w}^i||_2^2\eta_i}{2\,\sigma^2} - \frac{\lambda_2^2}{2\eta_i}\right\}$$
$$= \left(\frac{1}{\eta_i^3}\right)^{\frac{1}{2}} \exp\left\{-\frac{\lambda_2^2}{2}\left(\frac{\eta_i||\boldsymbol{w}^i||_2^2}{\lambda_2^2\,\sigma^2} + \frac{1}{\eta_i}\right)\right\}.$$

From this form we see the following result for distribution of η_i.

$$\eta_i = \frac{1}{\zeta_{\mathrm{i}}^2} \;\Big|\; \boldsymbol{Y}, \boldsymbol{W}, \boldsymbol{\tau}^2, \boldsymbol{\zeta}^{\boldsymbol{2}}_{\boldsymbol{(-i)}}, \sigma^2, \lambda_1^2, \lambda_2^2 \sim \textit{Inverse-Gaussian}\left(\sqrt{\frac{\lambda_2^2\,\sigma^2}{||\boldsymbol{w}^i||_2^2}}\;,\;\; \lambda_2^2\right).$$

5.5 Computation with Gibbs sampling

5.5.1 Results of full conditional distribution derivations

- $\text{vec}(\boldsymbol{W}^{(k)T})\big|\boldsymbol{Y}, \boldsymbol{W}^{(-k)}, \boldsymbol{\tau}^2, \boldsymbol{\zeta}^2, \sigma^2, \lambda_1^2, \lambda_2^2 \sim MVN_{m_k c}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$, for $k = 1, \ldots, K$

  $$\boldsymbol{\mu}_{\text{k}} = \boldsymbol{A}_k^{-1}\left(-\sum_{\ell=1}^{\text{n}}(\boldsymbol{x}_\ell^{(k)} \otimes I_c)(\boldsymbol{x}_\ell^{(-k)T} \otimes I_c)\text{vec}(\boldsymbol{W}^{(-k)T}) + \sum_{\ell=1}^{\text{n}}(\boldsymbol{x}_\ell^{(k)} \otimes I_c)\boldsymbol{y}_\ell\right)$$

  $$\boldsymbol{A}_k = \left(\sum_{\ell=1}^{\text{n}}(\boldsymbol{x}_\ell^{(k)} \otimes I_c)(\boldsymbol{x}_\ell^{(k)T} \otimes I_c) + \left(diag\left\{\frac{1}{\tau_{\text{k}}^2} + \frac{1}{\zeta_{\text{i}}^2}\right\}_{i \in \pi_k} \otimes I_c\right)\right)$$

  $$\boldsymbol{\Sigma}_k = \sigma^2\, \boldsymbol{A}_k^{-1}$$

- $\nu_k = \frac{1}{\tau_{\text{k}}^2}\left|\boldsymbol{Y}, \boldsymbol{W}, \tau_{(-k)}^2, \boldsymbol{\zeta}^2, \sigma^2, \lambda_1^2, \lambda_2^2 \sim \textit{Inverse-Gaussian}\left(\sqrt{\frac{\lambda_1^2\,\sigma^2}{||\boldsymbol{W}^{(k)}||_2^2}}\,, \lambda_1^2\right)\right.$,
  $k = 1, \ldots, K$

- $\eta_i = \frac{1}{\zeta_{\text{i}}^2}\left|\right. \boldsymbol{Y}, \boldsymbol{W}, \boldsymbol{\tau}^2, \zeta_{(-i)}^2, \sigma^2, \lambda_1^2, \lambda_2^2 \sim \textit{Inverse-Gaussian}\left(\sqrt{\frac{\lambda_2^2\,\sigma^2}{||\boldsymbol{w}^i||_2^2}}\,,\ \lambda_2^2\right),\ i = 1, \ldots, d$

- $\sigma^2\left|\boldsymbol{Y}, \boldsymbol{W}, \boldsymbol{\tau}^2, \boldsymbol{\zeta}^2, \lambda_1^2, \lambda_2^2 \sim\ \right. \textit{Inv} - \textit{Gamma}\,(a_\sigma^*, b_\sigma^*)$

  $$a_\sigma^* = \left(\tfrac{cn}{2} + \tfrac{dc}{2} + a_\sigma\right)$$

  $$b_\sigma^* = \left(\tfrac{1}{2}\sum_{\ell=1}^{n}||\boldsymbol{y}_\ell - \boldsymbol{W}^T\boldsymbol{x}_\ell||_2^2 + \tfrac{1}{2}\sum_{\text{i}=1}^{\text{d}}\frac{\sum_{\text{j}=1}^{\text{c}} w_{ij}^2}{\left(\frac{1}{\tau_{k(i)}^2} + \frac{1}{\zeta_i^2}\right)^{-1}} + b_\sigma\right)$$

5.5.2 Computation

The full conditional distributions allow for the implementation of Gibbs sampling for posterior sampling. Through extensive experimentation, we have discovered that the estimation of tuning parameters, λ_1^2 and λ_2^2, in this model is non-trivial for datasets where the number

of regression coefficients approaches the number of observations. For this reason, the next chapter is dedicated to discussing methods for the estimation of λ_1^2 and λ_2^2.

Chapter 6

Selection of tuning parameters

Park and Casella (2008) and Kyung et al. (2010) suggest two methods for tuning parameter estimation in a Bayesian framework: (i) putting λ^2's into the Gibbs sampler, and (ii) an empirical Bayes framework using marginal likelihoods. We adapt these methods to our model in Sections 6.1 and 6.2. For different settings of simulated data, preliminary results of these two approaches are displayed and discussed in Section 6.3. As a consequence of the results from the previous section, we study the shape of the marginal likelihood in Section 6.4, and finally discuss the use of cross-validation in Section 6.5.

6.1 Fully Bayesian model

Arguably, the easiest and fastest approach to estimate λ_1^2 and λ_2^2 is to assign them conjugate gamma prior distributions and enter them into the Gibbs iterations. Kyung et al. (2010) use the suggested gamma prior for a proper posterior (Park & Casella, 2008). They find this method to be the most attractive due to its efficiency, and also state that in the examples

that they investigated, the posterior means from the Gibbs iterations are very close to the marginal MLE.

6.1.1 Tuning parameter priors

Including λ_1^2 and λ_2^2 in the Gibbs iterations involves a simple extension to the hierarchical model. The priors are given by

- $\lambda_1^2 \sim Gamma(r_1, \delta_1), \qquad (r_1 > 0, \delta_1 > 0),$
- $\lambda_2^2 \sim Gamma(r_2, \delta_2), \qquad (r_2 > 0, \delta_2 > 0).$

6.1.2 Full conditionals

The full conditional distributions of λ_1^2 and λ_2^2 must be derived to allow for their inclusion in the Gibbs sampler.

Full conditional distribution of $\boldsymbol{\lambda_1^2}$

$$p(\lambda_1^2 \,|\boldsymbol{Y}, \boldsymbol{W}, \boldsymbol{\tau}^2, \boldsymbol{\zeta}^2, \sigma^2, \lambda_2^2) \propto \prod_{\mathrm{k}=1}^{\mathrm{K}} \left[\left(\frac{\lambda_1^2}{2} \right)^{\left(\frac{m_k c+1}{2}\right)} \exp\left\{ - \left(\frac{\lambda_1^2}{2} \right) \tau_{\mathrm{k}}^2 \right\} \right] (\lambda_1^2)^{r_1 - 1} \exp\left\{ -\delta_1 \, \lambda_1^2 \right\}$$

$$= \left(\frac{1}{2} \right)^{\frac{\sum_{\mathrm{k}=1}^{\mathrm{K}} m_k c + K}{2}} (\lambda_1^2)^{\frac{\sum_{\mathrm{k}=1}^{\mathrm{K}} m_k c + K}{2} + r_1 - 1} \exp\left\{ - \lambda_1^2 \, \frac{\sum_{\mathrm{k}=1}^{\mathrm{K}} \tau_{\mathrm{k}}^2}{2} - \lambda_1^2 \, \delta_1 \right\}.$$

Since $\left(\frac{1}{2}\right)^{\frac{\sum_{k=1}^{K} m_k c+K}{2}}$ can be factored out and $\sum_{k=1}^{K} m_k = d$, then

$$p(\lambda_1^2 \,|\boldsymbol{Y}, \boldsymbol{W}, \boldsymbol{\tau}^2, \boldsymbol{\zeta}^2, \sigma^2, \lambda_2^2) \propto (\lambda_1^2)^{\frac{dc+K}{2}+r_1-1} \exp\left\{-\lambda_1^2 \left(\frac{\sum_{k=1}^{K} \tau_k^2}{2} + \delta_1\right)\right\}.$$

Consequently,

$$\lambda_1^2 \,|\boldsymbol{Y}, \boldsymbol{W}, \boldsymbol{\tau}^2, \boldsymbol{\zeta}^2, \sigma^2, \lambda_2^2 \sim Gamma\left(\left(\frac{dc+K}{2} + r_1\right), \left(\frac{\sum_{k=1}^{K} \tau_k^2}{2} + \delta_1\right)\right).$$

Full conditional distribution of $\boldsymbol{\lambda_2^2}$

$$\begin{aligned} p(\lambda_2^2 \,|\boldsymbol{Y}, \boldsymbol{W}, \boldsymbol{\tau}^2, \boldsymbol{\zeta}^2, \sigma^2, \lambda_1^2) &\propto \prod_{i=1}^{d} \left[\left(\frac{\lambda_2^2}{2}\right)^{\left(\frac{c+1}{2}\right)} \exp\left\{-\left(\frac{\lambda_2^2}{2}\right)\zeta_i^2\right\}\right] (\lambda_2^2)^{r_2-1} \exp\left\{-\delta_2\,\lambda_2^2\right\} \\ &= \left(\frac{1}{2}\right)^{\frac{dc+d}{2}} (\lambda_2^2)^{\frac{dc+d}{2}+r_2-1} \exp\left\{-\lambda_2^2\,\frac{\sum_{i=1}^{d}\zeta_i^2}{2} - \lambda_2^2\,\delta_2\right\}. \end{aligned}$$

Factoring out $\left(\frac{1}{2}\right)^{\frac{dc+d}{2}}$ leaves

$$p(\lambda_2^2 \,|\boldsymbol{Y}, \boldsymbol{W}, \boldsymbol{\tau}^2, \boldsymbol{\zeta}^2, \sigma^2, \lambda_1^2) \propto (\lambda_1^2)^{\frac{d(c+1)}{2}+r_2-1} \exp\left\{-\lambda_2^2 \left(\frac{\sum_{i=1}^{d} \zeta_i^2}{2} + \delta_2\right)\right\}.$$

The full conditional of λ_2^2 is

$$\lambda_2^2 \,|\boldsymbol{Y}, \boldsymbol{W}, \boldsymbol{\tau}^2, \boldsymbol{\zeta}^2, \sigma^2, \lambda_1^2 \sim Gamma\left(\left(\frac{d(c+1)}{2} + r_2\right), \left(\frac{\sum_{i=1}^{d} \zeta_i^2}{2} + \delta_2\right)\right).$$

6.2 Empirical Bayes with Monte Carlo EM

The selection of tuning parameters using an Empirical Bayes approach is based on maximising the marginal likelihood of λ_1^2 and λ_2^2. Since the marginal likelihood, $p(\boldsymbol{Y} \mid \lambda_1^2, \lambda_2^2)$, is analytically intractable, a Monte Carlo EM algorithm is implemented to find the maximum likelihood estimates of λ_1^2 and λ_2^2.

6.2.1 Overview of Monte Carlo EM

We review the Expectation-Maximisation (EM) algorithm and subsequent Monte Carlo EM for a general model. This is based on work by Levine and Casella (2001).

Expectation-Maximisation (EM) algorithm

Let $\boldsymbol{y} = (y_1, \ldots, y_n)'$ denote data with distribution $f(\boldsymbol{y}|\boldsymbol{\Psi})$ characterised by the s-vector of parameters $\boldsymbol{\Psi}$. We wish to compute the marginal maximum likelihood estimate (MLE) of $\boldsymbol{\Psi}$. The EM algorithm can be used when the marginal MLE is easier to compute on the data augmented by a set of latent variables, $\boldsymbol{u} = (u_1, \ldots, u_q)'$. The augmented log-likelihood is given by $lnf(\boldsymbol{y}, \boldsymbol{u}|\boldsymbol{\Psi})$. The EM algorithm works on the augmented log-likelihood to obtain the marginal MLE of $\boldsymbol{\Psi}$ over the distribution $f(\boldsymbol{y}|\boldsymbol{\Psi})$, where it is assumed that $f(\boldsymbol{y}|\boldsymbol{\Psi}) = \int f(\boldsymbol{y}, \boldsymbol{u}|\boldsymbol{\Psi}) d\boldsymbol{u}$. The EM algorithm iterates through two steps. At the r^{th} iteration these steps are:

1. ***E-step:***

$$Q(\boldsymbol{\Psi}; \hat{\boldsymbol{\Psi}}^{(r-1)}) = E_{\hat{\boldsymbol{\Psi}}^{(r-1)}}(lnf(\boldsymbol{y}, \boldsymbol{u}|\boldsymbol{\Psi})|\boldsymbol{y})$$

   - the calculation of the expected complete-data log likelihood

2. ***M-step:***

$$\hat{\boldsymbol{\Psi}}^{(r)} = \arg\max_{\boldsymbol{\Psi}} \; Q(\boldsymbol{\Psi}|\hat{\boldsymbol{\Psi}}^{(r-1)})$$

- where the maximum value of $\boldsymbol{\Psi}$ is denoted by $\hat{\boldsymbol{\Psi}}^{(r)}$ and $\hat{\boldsymbol{\Psi}}^{(r-1)}$ denotes the maximum of $\boldsymbol{\Psi}$ at the $(r-1)^{th}$ iteration.

Monte Carlo EM algorithm

In situations where the E-step is analytically intractable, $Q(\boldsymbol{\Psi}; \hat{\boldsymbol{\Psi}}^{(r-1)})$ can be estimated with Monte Carlo integration.

$$E_{\hat{\boldsymbol{\Psi}}^{(r-1)}}(lnf(\boldsymbol{y}, \boldsymbol{u}|\boldsymbol{\Psi})|\boldsymbol{y}) = \int lnf(\boldsymbol{y}, \boldsymbol{u}|\boldsymbol{\Psi}) g(\boldsymbol{u}|\boldsymbol{y}, \hat{\boldsymbol{\Psi}}^{(r-1)}) du,$$

where $g(\boldsymbol{u}|\boldsymbol{y}, \boldsymbol{\Psi})$ is the conditional distribution of the latent variable, $\boldsymbol{u}$, given the data and $\boldsymbol{\Psi}$. A sample, $\boldsymbol{u}_1^{(r-1)}, \ldots, \boldsymbol{u}_m^{(r-1)}$, is obtained from the distribution $g(\boldsymbol{u}|\boldsymbol{y}, \hat{\boldsymbol{\Psi}}^{(r-1)})$. The expectation may then be estimated by the Monte Carlo sum,

$$Q_m(\boldsymbol{\Psi}; \hat{\boldsymbol{\Psi}}^{(r-1)}) = \frac{1}{m} \sum_{t=1}^{m} lnf(\boldsymbol{y}, \boldsymbol{u}_t^{(r-1)}|\boldsymbol{\Psi}).$$

For the Monte Carlo EM algorithm, the E-step of the EM algorithm is replaced by the estimated quantity above, and the M-step proceeds to maximise $Q_m(\boldsymbol{\Psi}; \hat{\boldsymbol{\Psi}}^{(r-1)})$ over $\boldsymbol{\Psi}$.

6.2.2 Monte Carlo EM for the estimation of λ_1^2 and λ_2^2

Setup and derivations

We implement a Monte Carlo EM (MCEM) algorithm to compute estimates, $\hat{\lambda_1^2}$ and $\hat{\lambda_2^2}$, that maximise the marginal likelihood, $p(\boldsymbol{Y}|\,\lambda_1^2, \lambda_2^2)$. The MCEM algorithm is set up by treating all other parameters in the model as *missing data*, in the traditional sense of the EM algorithm. In keeping with the notation of Section (6.2.1), $\boldsymbol{u} = (\boldsymbol{W}, \tau_1^2, \ldots, \tau_{\mathrm{K}}^2, \zeta_1^2, \ldots, \zeta_{\mathrm{d}}^2, \sigma^2)$ and $\boldsymbol{\Psi} = (\lambda_1^2, \lambda_2^2)$. The MCEM algorithm is used to find $\hat{\boldsymbol{\Psi}} = \arg\max_{\boldsymbol{\Psi}} p(\boldsymbol{Y}|\boldsymbol{\Psi})$.

The *complete-data* likelihood is given by

$$
\begin{aligned}
&p(\boldsymbol{Y},\boldsymbol{u}|\,\lambda_1^2, \lambda_2^2) \\
&\quad = p(\boldsymbol{Y}|\boldsymbol{W}, \sigma^2) p(\boldsymbol{W}|\,\sigma^2, \boldsymbol{\tau}^2, \boldsymbol{\zeta}^2) p(\boldsymbol{\tau}^2\,|\,\lambda_1^2) p(\boldsymbol{\zeta}^2\,|\,\lambda_2^2) p(\sigma^2\,|a_\sigma, b_\sigma) \\
&\quad = \prod_{\ell=1}^{\mathrm{n}} MVN_c\left(\boldsymbol{y}_\ell \middle| \boldsymbol{W}^T \boldsymbol{x}_\ell, \sigma^2\, I_c\right) \cdot \prod_{\mathrm{k}=1}^{\mathrm{K}} \prod_{\mathrm{i}\in\pi_{\mathrm{k}}} \prod_{\mathrm{j}=1}^{\mathrm{c}} N\left(w_{ij} \middle| 0, \sigma^2 \left(\frac{1}{\tau_k^2} + \frac{1}{\zeta_i^2}\right)^{-1}\right) \cdot \\
&\qquad \prod_{\mathrm{k}=1}^{\mathrm{K}} Gamma\left(\tau_k^2 \middle| \left(\frac{m_k c + 1}{2}\right), \left(\frac{\lambda_1^2}{2}\right)\right) \cdot \prod_{\mathrm{i}=1}^{\mathrm{d}} Gamma\left(\zeta_{\mathrm{i}}^2 \,\middle|\, \left(\frac{c+1}{2}\right), \left(\frac{\lambda_2^2}{2}\right)\right) \cdot \\
&\qquad Inv-Gamma(\sigma^2 \,|a_\sigma, b_\sigma) \\
&\quad \propto |\,\sigma^2\, I_c|^{-\frac{n}{2}} \exp\left\{-\frac{1}{2\,\sigma^2} \sum_{\ell=1}^{n} (\boldsymbol{y}_\ell - \boldsymbol{W}^T \boldsymbol{x}_\ell)^T (\boldsymbol{y}_\ell - \boldsymbol{W}^T \boldsymbol{x}_\ell)\right\} \\
&\qquad \prod_{\mathrm{k}=1}^{\mathrm{K}} \left[\prod_{i\in\pi_k} \left[\left(\sigma^2 \left(\frac{1}{\tau_k^2} + \frac{1}{\zeta_i^2}\right)^{-1}\right)^{-\frac{\mathrm{c}}{2}}\right] \exp\left\{-\sum_{\mathrm{i}\in\pi_{\mathrm{k}}} \left(\frac{\sum_{\mathrm{j}=1}^{\mathrm{c}} w_{ij}^2}{2\,\sigma^2 \left(\frac{1}{\tau_k^2} + \frac{1}{\zeta_i^2}\right)^{-1}}\right)\right\}\right] \\
&\qquad \prod_{\mathrm{k}=1}^{\mathrm{K}} \left[\frac{\left(\frac{\lambda_1^2}{2}\right)^{\left(\frac{m_k c+1}{2}\right)}}{\Gamma\left(\frac{m_k c+1}{2}\right)} (\tau_k^2)^{\left(\frac{m_k c+1}{2}\right)-1} \exp\left\{-\left(\frac{\lambda_1^2}{2}\right)\tau_k^2\right\}\right] \prod_{\mathrm{i}=1}^{\mathrm{d}} \left[\frac{\left(\frac{\lambda_2^2}{2}\right)^{\left(\frac{c+1}{2}\right)}}{\Gamma\left(\frac{c+1}{2}\right)} (\zeta_{\mathrm{i}}^2)^{\left(\frac{c+1}{2}\right)-1} \exp\left\{-\left(\frac{\lambda_2^2}{2}\right)\zeta_{\mathrm{i}}^2\right\}\right] \\
&\qquad \frac{b_\sigma^{a_\sigma}}{\Gamma(a_\sigma)} (\sigma^2)^{-a_\sigma - 1} \exp\left\{-\frac{b_\sigma}{\sigma^2}\right\}.
\end{aligned}
$$

After dropping terms that do not involve λ_1^2 and λ_2^2, the complete-data likelihood is proportional to

$$\prod_{\mathrm{k}=1}^{\mathrm{K}}\left[\frac{\left(\frac{\lambda_1^2}{2}\right)^{\left(\frac{m_k c+1}{2}\right)}}{\Gamma\left(\frac{m_k c+1}{2}\right)}(\tau_{\mathrm{k}}^2)^{\left(\frac{m_k c+1}{2}\right)-1}\exp\left\{-\left(\frac{\lambda_1^2}{2}\right)\tau_{\mathrm{k}}^2\right\}\right]\prod_{\mathrm{i}=1}^{\mathrm{d}}\left[\frac{\left(\frac{\lambda_2^2}{2}\right)^{\left(\frac{c+1}{2}\right)}}{\Gamma\left(\frac{c+1}{2}\right)}(\zeta_{\mathrm{i}}^2)^{\left(\frac{c+1}{2}\right)-1}\exp\left\{-\left(\frac{\lambda_2^2}{2}\right)\zeta_{\mathrm{i}}^2\right\}\right],$$

$$\propto (\lambda_1^2)^{\frac{dc+K}{2}}\exp\left\{-\lambda_1^2\frac{\sum_{\mathrm{k}=1}^{\mathrm{K}}\tau_{\mathrm{k}}^2}{2}\right\}(\lambda_2^2)^{\frac{dc+d}{2}}\exp\left\{-\lambda_2^2\frac{\sum_{\mathrm{i}=1}^{\mathrm{d}}\zeta_{\mathrm{i}}^2}{2}\right\}.$$

Accordingly, after dropping terms not involving λ_1^2 or λ_2^2, the complete-data log-likelihood is given by

$$\left(\frac{dc+K}{2}\right)log(\lambda_1^2)-\lambda_1^2\frac{\sum_{\mathrm{k}=1}^{\mathrm{K}}\tau_{\mathrm{k}}^2}{2}+\left(\frac{dc+d}{2}\right)log(\lambda_2^2)-\lambda_2^2\frac{\sum_{\mathrm{i}=1}^{\mathrm{d}}\zeta_{\mathrm{i}}^2}{2}.$$

At the r^{th} iteration of the Monte Carlo EM algorithm, the log-likelihood conditional on $\hat{\boldsymbol{\Psi}}^{(r-1)}=(\hat{\lambda_1^2}^{(r-1)},\hat{\lambda_2^2}^{(r-1)})$ and the data, $\boldsymbol{Y}$, is

$$\begin{aligned}
&Q(\boldsymbol{\Psi}|\hat{\boldsymbol{\Psi}}^{(r-1)})\\
&\quad=E\left[\left(\frac{dc+K}{2}\right)log(\lambda_1^2)-\lambda_1^2\frac{\sum_{\mathrm{k}=1}^{\mathrm{K}}\tau_{\mathrm{k}}^2}{2}+\left(\frac{dc+d}{2}\right)log(\lambda_2^2)-\lambda_2^2\frac{\sum_{\mathrm{i}=1}^{\mathrm{d}}\zeta_{\mathrm{i}}^2}{2}\ \Big|\hat{\boldsymbol{\Psi}}^{(r-1)},\ \boldsymbol{Y}\right]\\
&\quad=\left(\frac{dc+K}{2}\right)log(\lambda_1^2)-\frac{\lambda_1^2}{2}\sum_{\mathrm{k}=1}^{\mathrm{K}}E\left[\tau_{\mathrm{k}}^2\,|\hat{\boldsymbol{\Psi}}^{(r-1)},\boldsymbol{Y}\right]+\left(\frac{dc+d}{2}\right)log(\lambda_2^2)-\frac{\lambda_2^2}{2}\sum_{\mathrm{i}=1}^{\mathrm{d}}E\left[\zeta_{\mathrm{i}}^2\,|\hat{\boldsymbol{\Psi}}^{(r-1)},\boldsymbol{Y}\right].
\end{aligned}$$

The M-step of the algorithm determines $\boldsymbol{\Psi}=(\lambda_1^2,\lambda_2^2)$ that maximises $Q(\boldsymbol{\Psi}|\hat{\boldsymbol{\Psi}}^{(r-1)})$. We take the first derivative with respect to λ_1^2 and λ_2^2, respectively, set these to zero and solve for λ_1^2 and λ_2^2 to determine the marginal MLEs.

1) $\frac{\partial Q}{\partial\,\lambda_1^2}=\frac{dc+K}{2\,\lambda_1^2}-\frac{1}{2}\sum_{\mathrm{k}=1}^{\mathrm{K}}E\left[\tau_{\mathrm{k}}^2\,|\hat{\boldsymbol{\Psi}}^{(r-1)},\boldsymbol{Y}\right]$; $\hat{\lambda_1^2}^{(r)}=\frac{dc+K}{\sum_{\mathrm{k}=1}^{\mathrm{K}}E\left[\tau_{\mathrm{k}}^2\,|\hat{\boldsymbol{\Psi}}^{(r-1)},\boldsymbol{Y}\right]}$.

2) $\frac{\partial Q}{\partial \lambda_2^2} = \frac{dc+d}{2\,\lambda_2^2} - \frac{1}{2}\sum_{\mathrm{i}=1}^{\mathrm{d}} E\left[\zeta_{\mathrm{i}}^2 \,|\hat{\boldsymbol{\Psi}}^{(r-1)}, \boldsymbol{Y}\right]$; $\hat{\lambda_2^2}^{(r)} = \frac{dc+d}{\sum_{\mathrm{i}=1}^{\mathrm{d}} E\left[\zeta_{\mathrm{i}}^2 \,|\hat{\boldsymbol{\Psi}}^{(r-1)}, \boldsymbol{Y}\right]}$.

Summary of Monte Carlo EM for the estimation of λ_1^2 and λ_2^2

The Monte Carlo EM algorithm iterates through the E- and M-steps until the marginal MLE estimates from each iteration converge.

The E-step The E-step iterates through the Gibbs sampling algorithm to obtain a sample, $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_m$ from the distribution $g(\boldsymbol{u}|\hat{\boldsymbol{\Psi}}^{(r-1)}, \boldsymbol{Y})$, where $\boldsymbol{u} = (\boldsymbol{W}, \tau_1^2, \ldots, \tau_{\mathrm{K}}^2, \zeta_1^2, \ldots, \zeta_{\mathrm{d}}^2, \sigma^2)$ and $\boldsymbol{\Psi} = (\lambda_1^2, \lambda_2^2)$. The samples are used to evaluate

$$Q(\boldsymbol{\Psi}|\hat{\boldsymbol{\Psi}}^{(r-1)}) = E\left[\left(\frac{dc+K}{2}\right) log(\lambda_1^2) - \lambda_1^2\,\frac{\sum_{\mathrm{k}=1}^{\mathrm{K}} \tau_{\mathrm{k}}^2}{2} + \left(\frac{dc+d}{2}\right) log(\lambda_2^2) - \lambda_2^2\,\frac{\sum_{\mathrm{i}=1}^{\mathrm{d}} \zeta_{\mathrm{i}}^2}{2} \;\middle|\; \hat{\boldsymbol{\Psi}}^{(r-1)}, \boldsymbol{Y}\right].$$

Since $Q(\boldsymbol{\Psi}|\hat{\boldsymbol{\Psi}}^{(r-1)})$ is a linear function of $\tau_1^2, \ldots, \tau_{\mathrm{K}}^2$ and $\zeta_1^2, \ldots, \zeta_{\mathrm{d}}^2$,

$$Q(\boldsymbol{\Psi}|\hat{\boldsymbol{\Psi}}^{(r-1)}) = \left(\frac{dc+K}{2}\right) log(\lambda_1^2) - \frac{\lambda_1^2}{2}\sum_{\mathrm{k}=1}^{\mathrm{K}} E\left[\tau_{\mathrm{k}}^2 \,|\hat{\boldsymbol{\Psi}}^{(r-1)}, \boldsymbol{Y}\right] + \left(\frac{dc+d}{2}\right) log(\lambda_2^2) - \frac{\lambda_2^2}{2}\sum_{\mathrm{i}=1}^{\mathrm{d}} E\left[\zeta_{\mathrm{i}}^2 \,|\hat{\boldsymbol{\Psi}}^{(r-1)}, \boldsymbol{Y}\right]$$

and the expectations are evaluated with posterior means of $(\tau_1^2, \ldots, \tau_{\mathrm{K}}^2, \zeta_1^2, \ldots, \zeta_{\mathrm{d}}^2)$ from the Gibbs sampler.

The M-step The M-step finds a new estimate, $\hat{\boldsymbol{\Psi}}^{(r)}$, by maximising $Q(\boldsymbol{\Psi}|\hat{\boldsymbol{\Psi}}^{(r-1)})$ over $\boldsymbol{\Psi}$. As previously shown, $\hat{\boldsymbol{\Psi}}^{(r)} = (\hat{\lambda_1^2}^{(r)}, \hat{\lambda_2^2}^{(r)})$ is given by

$$\hat{\lambda_1^2}^{(r)} = \frac{dc+K}{\sum_{\mathrm{k}=1}^{\mathrm{K}} E\left[\tau_{\mathrm{k}}^2 \,|\hat{\boldsymbol{\Psi}}^{(r-1)}, \boldsymbol{Y}\right]} \quad \text{and} \quad \hat{\lambda_2^2}^{(r)} = \frac{dc+d}{\sum_{\mathrm{i}=1}^{\mathrm{d}} E\left[\zeta_{\mathrm{i}}^2 \,|\hat{\boldsymbol{\Psi}}^{(r-1)}, \boldsymbol{Y}\right]}.$$

6.3 Preliminary Gibbs sampling and empirical Bayes results

We begin investigating the behaviour of our MCMC algorithm by simulating data from the model, where the underlying true $\boldsymbol{W}$ is known. Data is simulated for different settings of d, the number of SNPs, c, the number of imaging phenotypes, and n, the number of observations. The model is fit using both the fully Bayesian and empirical Bayes framework for the estimation of tuning parameters, λ_1^2 and λ_2^2. We discover that the behaviour of model fitting changes drastically in two different settings.

6.3.1 Case 1: $d \ll n$

In situations where the number of SNPs included in the model is significantly smaller than the number of observations, both the full Gibbs and Empirical Bayes model perform well; Gibbs sampling λ_1^2 and λ_2^2 estimates converge to reasonable values and the Monte Carlo EM algorithm for λ_1^2 and λ_2^2 estimates converges. The resulting posterior means of $\boldsymbol{W}$ are close to the true $\boldsymbol{W}$ coefficient values. This behaviour is demonstrated by the results of a simulation with $d = 200,\ c = 5$, and $n = 500$.

Simulation settings and model fitting procedures

1. The Data Data is simulated for 500 subjects (n) as outlined in the following steps.

- A $\boldsymbol{W}$ matrix is simulated from its prior distribution in *Model B* with the following settings.

  - number of SNPs (d) = 200
  - SNPs are partitioned into 10 (K) groups
  - 10 groups of size 20
  - number of phenotypes (c) = 5
  - $\sigma^2 = \lambda_1^2 = \lambda_2^2 = 2$

- The SNP covariates used for data simulation are generated so that each SNP is generated from a discrete uniform distribution on $\{0, 1, 2\}$.
- The SNP covariates, $\boldsymbol{x}_\ell,\ \ell = 1, \ldots, n$, and the simulated $\boldsymbol{W}$ matrix are used to simulate c phenotypes for each subject based on the first level of the hierarchy in *Model B*. This generates $\boldsymbol{y}_\ell$, for $\ell = 1, \ldots, n$.

The fully Bayesian approach and MCEM algorithm are run on this set of simulated data.

2. Full Gibbs settings

- The parameters, $\sigma^2, \lambda_1^2, \lambda_2^2$, are assigned vague priors.
  - $\sigma^2 \sim Inv - Gamma(2, 1)$
  - $\lambda_1^2 \sim Gamma(0.00001, 0.00001)$
  - $\lambda_2^2 \sim Gamma(0.00001, 0.00001)$
- The Gibbs sampler is run for $20,000$ iterations and the first $10,000$ iterations are thrown out.

3. MCEM algorithm settings

- The prior assigned to σ^2 for the Gibbs sampling portion of the MCEM algorithm is the same vague prior as in the fully Bayesian approach, $\sigma^2 \sim Inv - Gamma(2, 1)$.
- The MCEM algorithm runs through 50 iterations.
  - The first *E-step* runs through 1500 iterations of the Gibbs sampler, where only the last 500 are used for computing expectations.
  - Subsequent *E-step*'s are reduced to including 600 iterations of Gibbs sampling, where starting values for parameters are set as the posterior means from the previous *E-step*. The last 500 iterations are used for computing expectations.

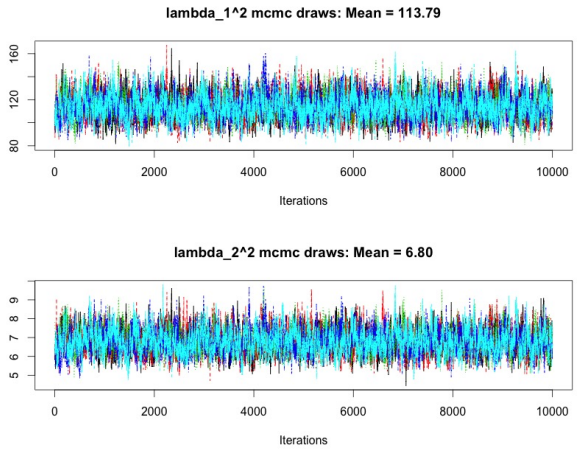


Figure 6.1: Case 1 Full Gibbs estimates

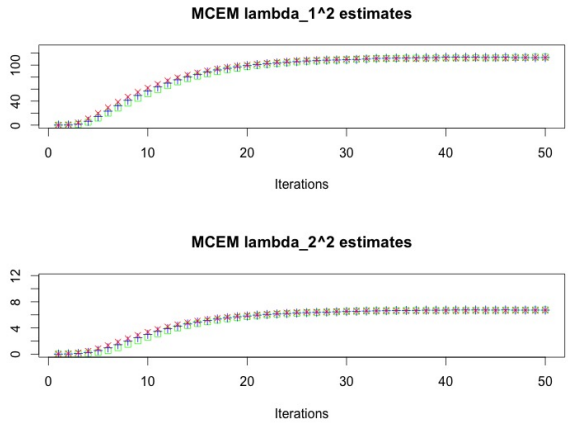


Figure 6.2: Case 1 MCEM estimates

Results

Figure 6.1 shows draws of λ_1^2 and λ_2^2 from the second half of five MCMC chains. Posterior means of λ_1^2 and λ_2^2 are 113.79 and 6.80 respectively. Figure 6.2 shows λ_1^2 and λ_2^2 estimates from three MCEM runs with different colours and characters representing different starting

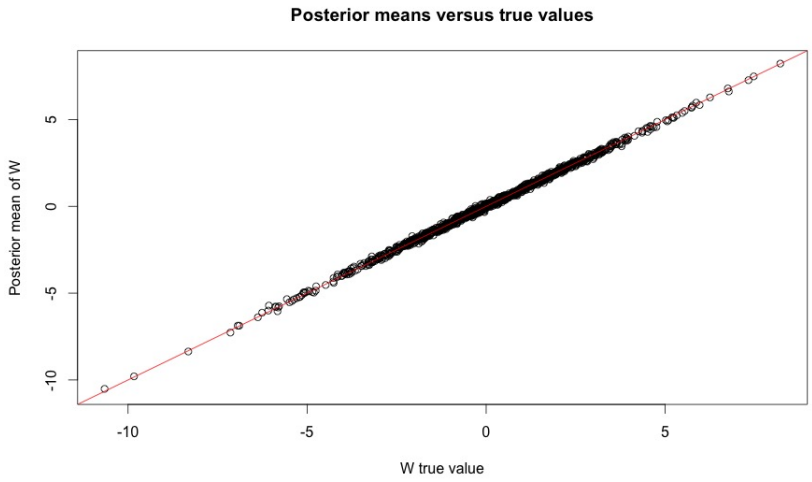


Figure 6.3: Case 1 posterior means

values for λ_1^2 and λ_2^2. The estimates converge to approximately 111.3 and 6.6. Finally, Figure 6.3 shows the posterior means of $\boldsymbol{W}$ from the full Gibbs model plotted against the true $\boldsymbol{W}$ values.

6.3.2 Case 2: $d \approx$ or $\geq n$

In situations where the number of SNPs included in the model is approximately equal to or greater than the number of observations problems arise; Gibbs sampling λ_1^2 and λ_2^2 estimates converge to unreasonably large values and the Monte Carlo EM algorithm for λ_1^2 and λ_2^2 estimates diverges. The resulting posterior means of $\boldsymbol{W}$ are over-shrunk and are very poor estimates. This behaviour is demonstrated by the results of a simulation with $d = 510, c = 5$, and $n = 500$.

Simulation settings and model fitting procedures

1. The Data Data is simulated for 500 subjects (n) as outlined in the following steps.

- A $\boldsymbol{W}$ matrix is simulated from its prior distribution in *Model B* with the following settings.
  - number of SNPs (d) = 510
  - SNPs are partitioned into 43 (K) groups
  - 20 groups of size 20, 9 groups of size 10, 6 groups of size 2, 8 groups of size 1
  - number of phenotypes (c) = 5
  - $\sigma^2 = \lambda_1^2 = \lambda_2^2 = 2$
- The SNP covariates used for data simulation are generated so that each SNP is generated from a discrete uniform distribution on $\{0, 1, 2\}$.
- The SNP covariates, $\boldsymbol{x}_\ell,\ \ell = 1, \ldots, n$, and the simulated $\boldsymbol{W}$ matrix are used to simulate c phenotypes for each subject based on the first level of the hierarchy in *Model B*. This generates $\boldsymbol{y}_\ell$, for $\ell = 1, \ldots, n$.

We run the fully Bayesian approach and MCEM algorithm on this set of simulated data.

2. Full Gibbs and MCEM algorithm settings The remaining model fitting procedures are the same as those described for **Case 1** in Section 6.3.1.

Results

Figure 6.4 shows draws of λ_1^2 and λ_2^2; both have posterior means much greater than one million. Figure 6.5 shows λ_1^2 and λ_2^2 estimates from MCEM, which diverge to infinity. From

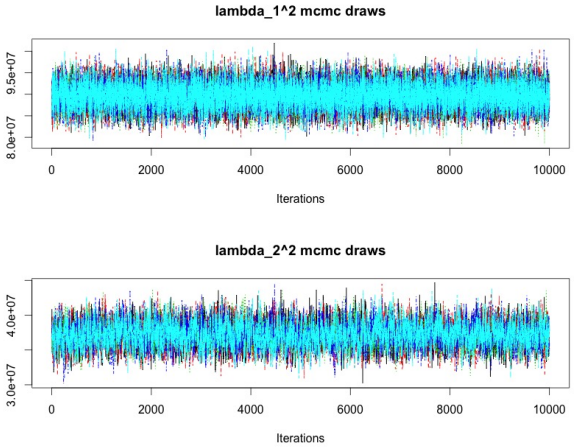


Figure 6.4: Case 2 Full Gibbs estimates

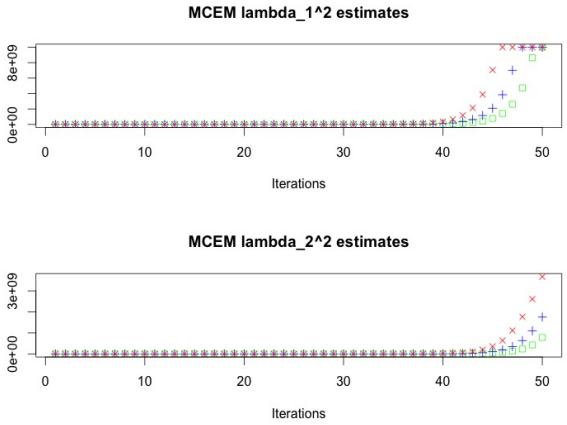


Figure 6.5: Case 2 MCEM estimates

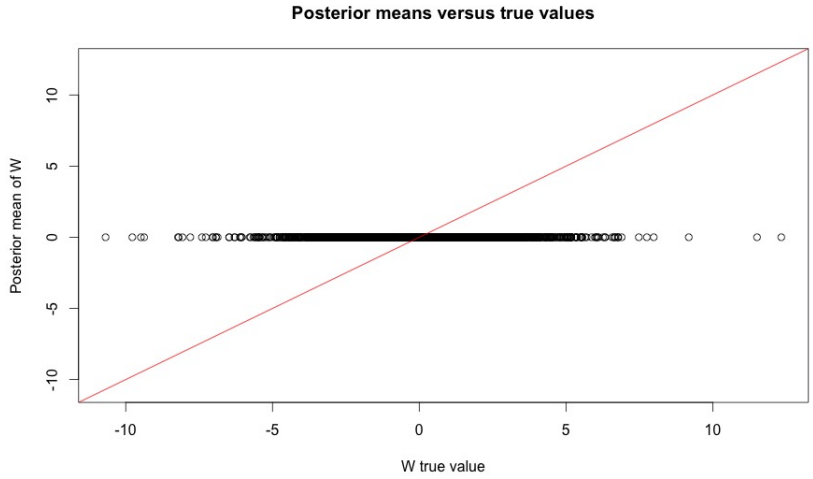


Figure 6.6: Case 2 posterior means

Figure 6.6, the posterior means of $\boldsymbol{W}$ from the full Gibbs model plotted against the true $\boldsymbol{W}$ values, we see that coefficient estimates have been over-shrunk by the large λ_1^2 and λ_2^2 values.

6.3.3 Discussion

It seems that with a large number of SNPs, relative to the number of observations, choosing the tuning parameters based on the posterior distributions or the marginal likelihood leads

to over shrinking of coefficients.

Aside from the full Gibbs model and MCEM for estimation of the λ^2's, we note that when λ_1^2 and λ_2^2 are fixed at their true values, the MCMC algorithm performs well in cases where $d \geq n$. Figure 6.7 shows the results of running the Gibbs sampling algorithm, with parameters λ_1^2 and λ_2^2 fixed at their true values, on the same dataset described and investigated in **Case 2**, with $d = 510, c = 5$, and $n = 500$.

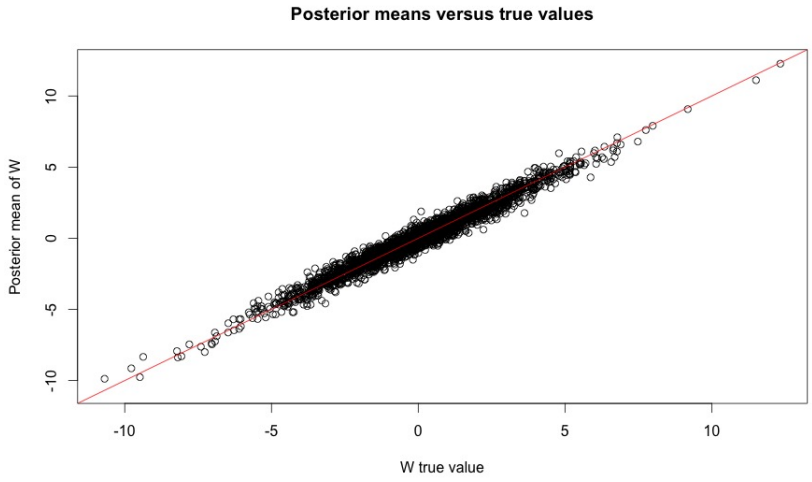


Figure 6.7: $\boldsymbol{W}$ Posterior means from fixed λ_1^2 and λ_2^2

6.4 Studying the marginal likelihood

In order to better understand the model we study the shape of the marginal likelihood, $p(\boldsymbol{Y} \mid \lambda_1^2, \lambda_2^2)$. Since parts of the integration involved in solving an expression for $p(\boldsymbol{Y} \mid \lambda_1^2, \lambda_2^2)$ are analytically intractable, we derive an approximation to the marginal likelihood and study its shape.

6.4.1 Derivation

$$p(\boldsymbol{Y}|\,\lambda_1^2,\lambda_2^2) = \int p\,(\boldsymbol{Y},\boldsymbol{W},\sigma^2,\boldsymbol{\tau}^2,\boldsymbol{\zeta}^2\,|\,\lambda_1^2,\lambda_2^2)\,d\boldsymbol{W}\,d\,\sigma^2\,d\,\boldsymbol{\tau}^2\,d\,\boldsymbol{\zeta}^2 \tag{6.1}$$

- Let $\boldsymbol{y} = (y_{11},\ldots,y_{n1},\ldots,y_{1c},\ldots,y_{nc})^T$, a vector of observations blocked by imaging phenotype.
- Let $\boldsymbol{X} = (\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n)^T$, an $n \times d$ matrix of SNP data.
- Let $\widetilde{\boldsymbol{w}} = (w_{11},\ldots,w_{d1},\ldots,w_{1c},\ldots,w_{dc})^T$, a vector of SNP regression coefficients blocked by imaging phenotype.

Under the assumed hierarchical *Model B* we have,

$$\widetilde{\boldsymbol{w}}|\,\boldsymbol{\tau}^2,\boldsymbol{\zeta}^2,\sigma^2 \sim MVN(0,\boldsymbol{\Sigma}_{\widetilde{\boldsymbol{w}}}),$$

$$\text{where } \boldsymbol{\Sigma}_{\widetilde{\boldsymbol{w}}} = \sigma^2\,I_c \otimes Diag\left\{\left(\frac{1}{\zeta_{\mathrm{i}}^2}+\frac{1}{\tau_{k(i)}^2}\right)^{-1}, i=1,\ldots,d\right\}.$$

Under the model we also have

$$\boldsymbol{y} \sim MVN\left((I_c \otimes \boldsymbol{X})\widetilde{\boldsymbol{w}},\sigma^2\,I_{cn}\right).$$

Note: If we have $\boldsymbol{x}$ and $\boldsymbol{y}$ vectors such that $\boldsymbol{x} \sim MVN(\boldsymbol{\mu},\Lambda)$ and $\boldsymbol{y}|\boldsymbol{x} \sim MVN(A\boldsymbol{x}+b,S)$, where Λ and S are covariance matrices, it can be shown that we obtain the marginal likelihood of $\boldsymbol{y}$ as follows: $\boldsymbol{y} \sim N(A\boldsymbol{\mu}+b,S+A\Lambda A^T)$ (Bishop, 2006).

From this property of the Gaussian distribution, $\widetilde{\boldsymbol{w}}$ is marginalised out of (6.1) to give

$$\boldsymbol{y}|\,\boldsymbol{\tau}^2,\boldsymbol{\zeta}^2,\sigma^2 \sim MVN(0,(I_c \otimes \boldsymbol{X})\boldsymbol{\Sigma}_{\widetilde{\boldsymbol{w}}}(I_c \otimes \boldsymbol{X}^T)+\sigma^2\,I_{cn}).$$

The marginal likelihood is then expressed as

$$\int p(\boldsymbol{y}, \mid \sigma^2, \boldsymbol{\tau}^2, \boldsymbol{\zeta}^2) p(\sigma^2)\; d\sigma^2 p(\boldsymbol{\tau}^2 \mid \lambda_1^2) p(\boldsymbol{\zeta}^2 \mid \lambda_2^2)\; d\,\boldsymbol{\tau}^2\, d\,\boldsymbol{\zeta}^2\,.$$

This expression is rewritten as

$$p(\boldsymbol{y}|\, \lambda_1^2, \lambda_2^2) = \int \underbrace{\int_0^\infty p(\boldsymbol{y}, \mid \sigma^2, \boldsymbol{\tau}^2, \boldsymbol{\zeta}^2) p(\sigma^2)\; d\,\sigma^2}_{\bigstar}\, p(\boldsymbol{\tau}^2 \mid \lambda_1^2) p(\boldsymbol{\zeta}^2 \mid \lambda_2^2)\; d\,\boldsymbol{\tau}^2\, d\,\boldsymbol{\zeta}^2\,.$$

We first investigate the integration in $\bigstar$.

$$\int_0^\infty p(\boldsymbol{y}, \mid \sigma^2, \boldsymbol{\tau}^2, \boldsymbol{\zeta}^2) p(\sigma^2) d\,\sigma^2$$

$$\begin{aligned}
&= \int_0^\infty (2\pi)^{-\frac{nc}{2}} \left|(I_c \otimes \boldsymbol{X}) \boldsymbol{\Sigma}_{\widetilde{\boldsymbol{w}}} (I_c \otimes \boldsymbol{X}^T) + \sigma^2\, I_{cn}\right|^{-\frac{1}{2}} \times \\
&\quad exp\left\{ -\frac{1}{2} \boldsymbol{y}^T \left\{ (I_c \otimes \boldsymbol{X}) \boldsymbol{\Sigma}_{\widetilde{\boldsymbol{w}}} (I_c \otimes \boldsymbol{X}^T) + \sigma^2\, I_{cn} \right\}^{-1} \boldsymbol{y} \right\} \times \frac{b_\sigma^{a_\sigma}}{\Gamma(a_\sigma)} (\sigma^2)^{-(a_\sigma+1)}\; exp\{-\frac{b_\sigma}{\sigma^2}\} d\,\sigma^2
\end{aligned}$$

$$\begin{aligned}
&= (2\pi)^{-\frac{nc}{2}} \left| (I_c \otimes \boldsymbol{X}) \left(\mathrm{I_c} \otimes \mathrm{Diag} \left\{ \left(\frac{1}{\zeta_{\mathrm{i}}^2} + \frac{1}{\tau_{\mathrm{k(i)}}^2} \right)^{\text{-1}} \right\} \right) (I_c \otimes \boldsymbol{X}^T) + I_{cn} \right|^{-\frac{1}{2}} \times \frac{b_\sigma^{a_\sigma}}{\Gamma(a_\sigma)} \times \\
&\quad \int_0^\infty (\sigma^2)^{-(\frac{nc}{2}+a_\sigma+1)} \\
&\quad \underbrace{exp\left\{ -\frac{1}{\sigma^2} \left(b_\sigma + \frac{1}{2} \boldsymbol{y}^T \left[(I_c \otimes \boldsymbol{X}) \left(\mathrm{I_c} \otimes \mathrm{Diag} \left\{ \left(\frac{1}{\zeta_{\mathrm{i}}^2} + \frac{1}{\tau_{\mathrm{k(i)}}^2} \right)^{\text{-1}} \right\} \right) (I_c \otimes \boldsymbol{X}^T) + I_{cn} \right]^{-1} \boldsymbol{y} \right) \right\} d\,\sigma^2}_{\text{Kernel of Inv-Gamma}}
\end{aligned}$$

We solve the integral in the above expression to get

$$p(\boldsymbol{y}|\,\boldsymbol{\tau}^2,\boldsymbol{\zeta}^2) =$$
$$\left[(2\pi)^{-\frac{nc}{2}}\ b_\sigma^{a_\sigma}\ \frac{\Gamma(\frac{nc}{2}+a_\sigma)}{\Gamma(a_\sigma)}\right]\times\left|(I_c\otimes\boldsymbol{X})\left(\mathrm{I_c}\otimes\mathrm{Diag}\left\{\left(\frac{1}{\zeta_{\mathrm{i}}^2}+\frac{1}{\tau_{\mathrm{k(i)}}^2}\right)^{\text{-1}}\right\}\right)(I_c\otimes\boldsymbol{X}^T)+I_{cn}\right|^{-\frac{1}{2}}\times$$
$$\left(b_\sigma+\frac{1}{2}\boldsymbol{y}^T\left[(I_c\otimes\boldsymbol{X})\left(\mathrm{I_c}\otimes\mathrm{Diag}\left\{\left(\frac{1}{\zeta_{\mathrm{i}}^2}+\frac{1}{\tau_{\mathrm{k(i)}}^2}\right)^{\text{-1}}\right\}\right)(I_c\otimes\boldsymbol{X}^T)+I_{cn}\right]^{-1}\boldsymbol{y}\right)^{-(\frac{nc}{2}+a_\sigma)}. \tag{6.2}$$

Then,

$$p(\boldsymbol{y}|\,\lambda_1^2,\lambda_2^2)=\int p(\boldsymbol{y}|\,\boldsymbol{\tau}^2,\boldsymbol{\zeta}^2)p(\boldsymbol{\tau}^2\mid\lambda_1^2)p(\boldsymbol{\zeta}^2\mid\lambda_2^2)d\,\boldsymbol{\tau}^2\,d\,\boldsymbol{\zeta}^2=E_{\boldsymbol{\tau}^2,\boldsymbol{\zeta}^2}\left[p(\boldsymbol{y}|\,\boldsymbol{\tau}^2,\boldsymbol{\zeta}^2)\right],$$

where $p(\boldsymbol{y}|\,\boldsymbol{\tau}^2,\boldsymbol{\zeta}^2)$ in given by (6.2). The remaining integration is not analytically tractable, and we use a plug-in approximation,

$$p(\boldsymbol{y}|\,\lambda_1^2,\lambda_2^2)=E_{\boldsymbol{\tau}^2,\boldsymbol{\zeta}^2}\left[p(\boldsymbol{y}|\,\boldsymbol{\tau}^2,\boldsymbol{\zeta}^2)\right]\approx p(\boldsymbol{y}|E[\boldsymbol{\tau}^2],E[\boldsymbol{\zeta}^2]),$$

$$E[\tau_{\mathrm{k}}^2]=\frac{m_kc+1}{\lambda_1^2},\quad k=1,\ldots,K,\quad\text{and}\quad E[\zeta_{\mathrm{i}}^2]=\frac{c+1}{\lambda_2^2},\quad i=1,\ldots,d,$$

where we have assumed the distributions form *Model B* for $\boldsymbol{\tau}^2$ and $\boldsymbol{\zeta}^2$. This gives the following closed form approximation:

$$
\begin{aligned}
&p(\boldsymbol{y}|\,\lambda_1^2,\ \lambda_2^2) \approx \\
&(2\pi)^{-\frac{nc}{2}}\ b_\sigma^{a_\sigma}\ \frac{\Gamma(\frac{nc}{2}+a_\sigma)}{\Gamma(a_\sigma)} \times \left| (\mathrm{I_c}\otimes \mathrm{X}) \left(\mathrm{I_c}\otimes \mathrm{Diag}\left\{ \left(\frac{\lambda_2^2}{\mathrm{c}+1} + \frac{\lambda_1^2}{\mathrm{m_{k(i)}c}+1} \right)^{\text{-1}}, \mathrm{i}=1,\ldots,\mathrm{d} \right\} \right) (\mathrm{I_c}\otimes \mathrm{X^T}) + \mathrm{I_{cn}}\ ) \right|^{-\frac{1}{2}} \times \\
&\left(b_\sigma + \frac{1}{2}\boldsymbol{y}^T \left[(\mathrm{I_c}\otimes \mathrm{X}) \left(\mathrm{I_c}\otimes \mathrm{Diag}\left\{ \left(\frac{\lambda_2^2}{\mathrm{c}+1} + \frac{\lambda_1^2}{\mathrm{m_{k(i)}c}+1} \right)^{\text{-1}}, \mathrm{i}=1,\ldots,\mathrm{d} \right\} \right) (\mathrm{I_c}\otimes \mathrm{X^T}) + \mathrm{I_{cn}}\ ) \right]^{-1} \boldsymbol{y} \right)^{-(\frac{nc}{2}+a_\sigma)}.
\end{aligned}
\tag{6.3}
$$

6.4.2 Approximation plots

Since we are primarily concerned with investigating how the shape of the marginal likelihood changes as it is plotted over different values of λ_1^2 and λ_2^2, terms of (6.3) that do not involve λ_1^2 or λ_2^2 are dropped from the approximation. We evaluate the natural log of

$$
\begin{aligned}
&f(\lambda_1^2,\lambda_2^2) = \\
&\quad \left| (\mathrm{I_c}\otimes \mathrm{X}) \left(\mathrm{I_c}\otimes \mathrm{Diag}\left\{ \left(\frac{\lambda_2^2}{\mathrm{c}+1} + \frac{\lambda_1^2}{\mathrm{m_{k(i)}c}+1} \right)^{\text{-1}}, \mathrm{i}=1,\ldots,\mathrm{d} \right\} \right) (\mathrm{I_c}\otimes \mathrm{X^T}) + \mathrm{I_{cn}}\ ) \right|^{-\frac{1}{2}} \times \\
&\left(b_\sigma + \frac{1}{2}\boldsymbol{y}^T \left[(\mathrm{I_c}\otimes \mathrm{X}) \left(\mathrm{I_c}\otimes \mathrm{Diag}\left\{ \left(\frac{\lambda_2^2}{\mathrm{c}+1} + \frac{\lambda_1^2}{\mathrm{m_{k(i)}c}+1} \right)^{\text{-1}}, \mathrm{i}=1,\ldots,\mathrm{d} \right\} \right) (\mathrm{I_c}\otimes \mathrm{X^T}) + \mathrm{I_{cn}}\ ) \right]^{-1} \boldsymbol{y} \right)^{-(\frac{nc}{2}+a_\sigma)}
\end{aligned}
\tag{6.4}
$$

over a grid of $(\lambda_1^2, \lambda_2^2)$ values to construct plots for different sets of simulated data. We show plots of $log\,(f(\lambda_1^2, \lambda_2^2))$ for two different simulated datasets, which demonstrate contrasting shapes in the marginal likelihood approximation.

Dataset 1

Data description The simulated data plotted here is the same data used for **Case 1** in Section 6.3.1. This simulation setting has $d = 200,\ c = 5$, and $n = 500$.

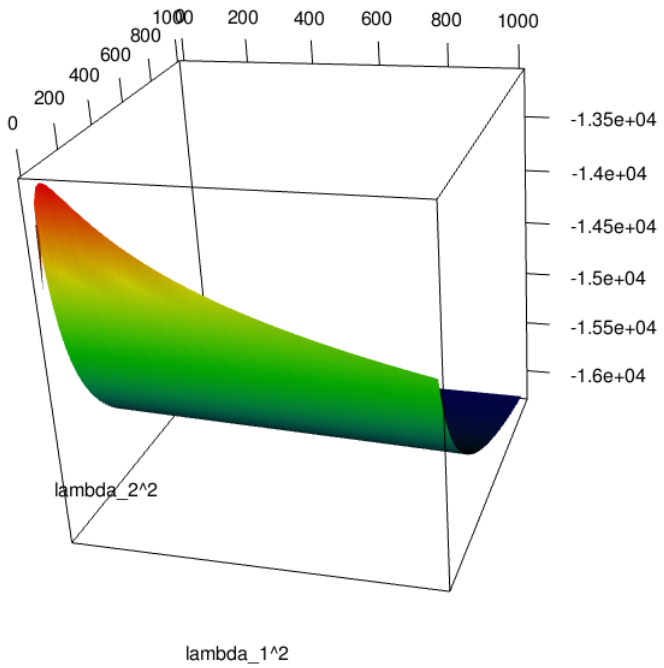


Figure 6.8: Dataset 1 ML approximation shape

Marginal likelihood approximation shape Figure 6.8 shows the shape of the approximation. The natural log of $f(\lambda_1^2, \lambda_2^2)$ is evaluated over a 100×100 grid with λ_1^2 and λ_2^2 values ranging from 0.1 to 1000. This plot shows a peak near the origin, and the maximum point is found at $\lambda_1^2 = 30.4$, $\lambda_2^2 = 0.1$. When these values are plugged back into the Gibbs sampler as fixed $\hat{\lambda}_1^2$ and $\hat{\lambda}_2^2$, the posterior means of $\boldsymbol{W}$ are close to the true values.

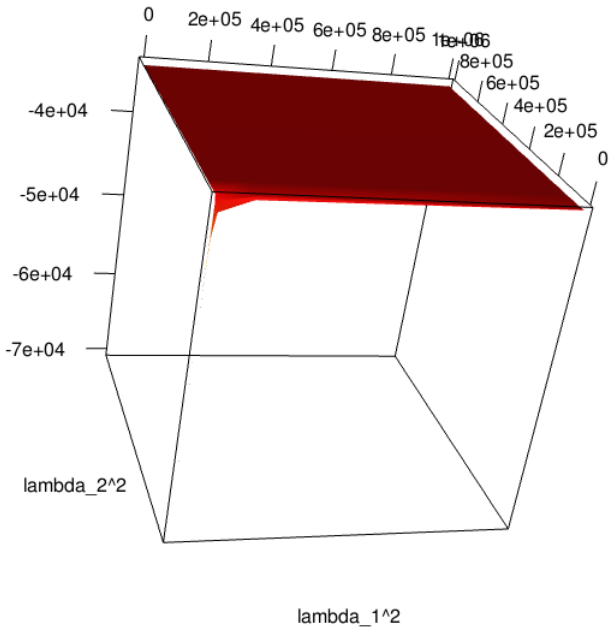


Figure 6.9: Dataset 2 ML approximation shape

Dataset 2

Data description This data is simulated to resemble what we suspect the real data looks like, in particular it is simulated with weak effects. The SNP covariates used for data simulation are real genetic data on 632 subjects over 486 SNPs belonging to 33 different genes. This data is described in Section 2.1. A $\boldsymbol{W}$ matrix is simulated from its prior distribution in *Model B* with $d = 486$ and $c = 12$, where the grouping structure of SNPs is determined by the real genetic data, and true values of σ^2, λ_2^2 and λ_1^2 are 2. To induce weak effects in the data, each element of the simulated $\boldsymbol{W}$ matrix is divided by a factor of 100. This new $\boldsymbol{W}$ is used, in combination with the genetic data, to simulate the phenotypes for 632 subjects based on the first level of *Model B*.

Marginal likelihood approximation shape Figure 6.9 shows the shape of the approximation for this dataset simulated with weak effects and real genetic data. This shape

is generated by evaluating the natural log of $f(\lambda_1^2, \lambda_2^2)$ over a grid of λ_1^2 and λ_2^2 values in $(10^{-4}, 10^{-3}, \ldots, 10^5, 10^6)$. As the surface of this plot is flat, a flatness extending out towards the boundaries of λ_1^2 and λ_2^2 values, it is not surprising that, for this dataset, choosing tuning parameters based on this function leads to large λ_1^2 and λ_2^2 estimates. Indeed, the maximum point on this plot is at $\lambda_1^2 = 10^5$, $\lambda_2^2 = 10^4$.

6.5 Cross validation and WAIC

For Bayesian lasso and related hierarchical models Park and Casella (2008) and Kyung et al. (2010) found that "*putting λ into the Gibbs sampler seems as effective as choosing it by cross-validation*". For the model we have developed, when the number of SNPs is large and/or there are weak effects, we find empirically that cross-validation avoids some of the observed problems with full Bayesian and maximum marginal likelihood with MCEM choice of the tuning parameters. Combining Gibbs sampling with cross-validation over a two-dimensional grid of tuning parameters is, however, computationally intensive.

We, therefore, use WAIC (Watanabe, 2010), which does not require any data splitting for its computation and can be viewed as an approximation to leave-one-out cross-validation (Gelman, Hwang, & Vehtari, 2014). The form of the likelihood chosen for computation of WAIC is based on the fact that we are interested in predicting $\boldsymbol{Y}$ from $\boldsymbol{W}$, hence the primary parameter of interest is $\boldsymbol{W}$. Consequently, the form of the selected likelihood is

$p(\boldsymbol{Y}|\boldsymbol{W}, \sigma^2)$, which is defined at the first level of the hierarchy in *Model B*. From this,

$$WAIC = -2\sum_{\ell=1}^{n} \log E_{\boldsymbol{W},\sigma^2}[p(\boldsymbol{y}_\ell|\boldsymbol{W}, \sigma^2)|\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n] + 2\sum_{\ell=1}^{n} Var_{\boldsymbol{W},\sigma^2}[\log p(\boldsymbol{y}_\ell|\boldsymbol{W}, \sigma^2)|\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n], \quad (6.5)$$

where the posterior means and variances are approximated based on the output of the Gibbs sampler. The Gibbs-WAIC method applied in the next chapter runs Gibbs samplers in parallel over a two-dimensional grid of λ_1^2, λ_2^2 values and chooses the tuning parameters and model that minimises WAIC.

Chapter 7

Experimental results

A simulation study is performed to assess the properties of our Bayesian-WAIC method, and also to compare these results to those of the Wang et al. method. Subsequently, in Section 7.2, we apply both methods to the dataset obtained from the ADNI database.

7.1 Simulation study

7.1.1 Setup and method

The Data

The SNP covariates used for data simulation come from the ADNI database. This includes genetic data on 632 subjects over 486 SNPs belonging to 33 different genes, as described in Section 2.1.

A $\mathbf{W}$ matrix is simulated from its prior distribution in *Model B* with the following settings.

- number of SNPs (d) = 486
- SNPs are partitioned into 33 (K) genes
- number of phenotypes (c) = 12
- $\sigma^2 = \lambda_1^2 = \lambda_2^2 = 2$

Sparsity is then introduced to $\mathbf{W}$ by setting all but 50 rows to zero. Only the following rows are left at their simulated values.

- rows corresponding to 5 genes of SNP sizes 14, 10, 6, 4, 1 (35 SNPs)
- rows corresponding to 15 other randomly selected SNPs

The genetic data and sparse $\mathbf{W}$ matrix are used to simulate 100 sets of response variables.

Method

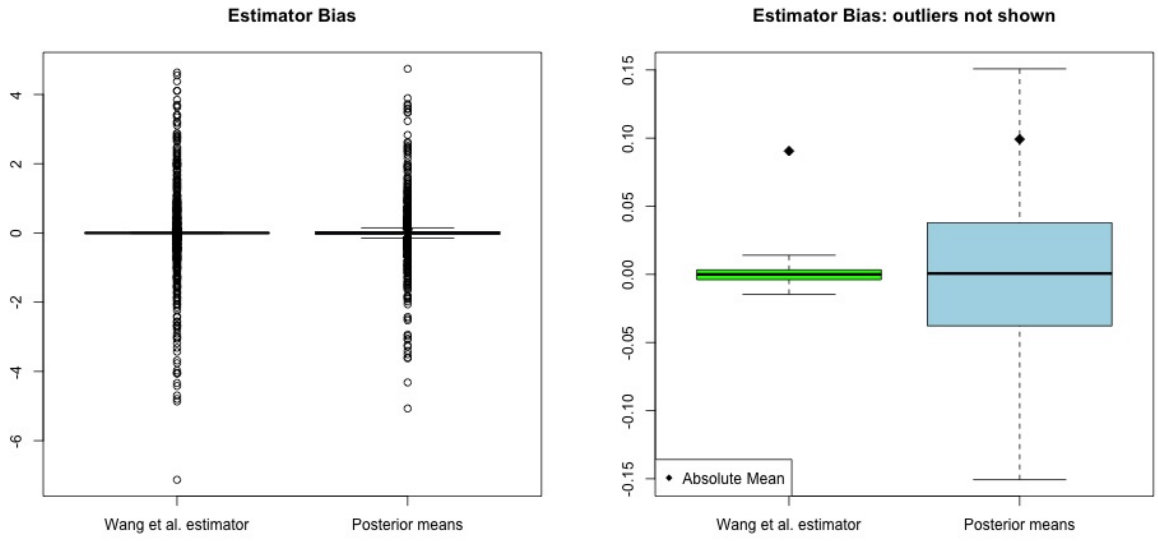
The Wang et al. method and our Gibbs-WAIC Bayesian method are applied to each of the 100 datasets. The estimator of $\boldsymbol{W}$ based on the Wang et al. method is computed with the algorithm provided by the authors, where the tuning parameters are chosen via 5-fold cross validation, as outlined in Chapter 4. Under the Bayesian framework, σ^2 is assigned a prior of $\sigma^2 \sim Inv - Gamma(3, 1)$. Bayesian model fitting is performed by fitting the model with fixed λ_1^2, λ_2^2 values over a two-dimensional grid of $\{0.01, 0.1, 1, 10, 100\}^2$ for a total of 25 MCMC runs in each dataset. Each MCMC run includes $10,000$ iterations of the Gibbs sampler, where the first 5000 iterations are treated as the *burn-in* period. WAIC is computed with posterior draws after the *burn-in* period. The model with tuning parameter values minimising WAIC is selected.

7.1.2 Simulation results

The Wang et al. method selects tuning parameter values, γ_1 and γ_2 from (4.1), of $\gamma_1 = 10^{-4}$ and $\gamma_2 = 10^2$ for all 100 simulations. The Gibbs-WAIC approach selects $\lambda_1^2 = 10$ and $\lambda_2^2 = 10$ for 97 of the 100 simulations. For the remaining 3 simulations, values of $\lambda_1^2 = 10^{-2}$ and $\lambda_2^2 = 10^2$ are selected with the Gibbs-WAIC approach.

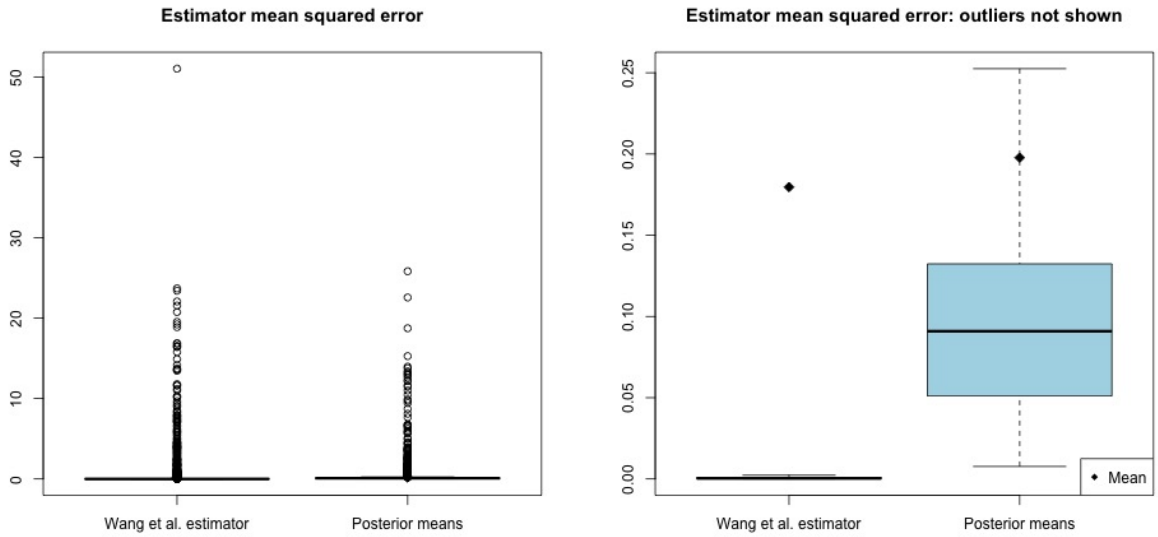
The estimated $\boldsymbol{W}$ matrix consists of 5832 regression coefficients. For this reason, we use boxplots to display properties of the Wang et al. estimator and the posterior means of $\boldsymbol{W}$ of our Bayesian model. Figure 7.1a shows boxplots of the bias from the two methods. Bias of the Wang et al. estimator has a range of $[-7.13, 4.63]$, while the bias of posterior means has range $[-5.07, 4.77]$. Boxplots of estimator bias without outliers are shown in Figure 7.1b. Here we see that the bias of the Wang et al. estimator is more concentrated around zero than the bias of posterior means. The two methods do, nonetheless, have similar mean absolute bias; the mean absolute bias from the Wang et al. estimator is 0.0905, while the mean absolute bias of posteriors means is 0.0992. Boxplots of the mean squared error (MSE) from the Wang et al. estimator and posteriors means are displayed in Figure 7.2, where 7.2a includes outliers and Figure 7.2b does not. Figure 7.2b shows that after excluding outlier mean squared errors of coefficient estimates, the Wang et al. estimator is more stable than the posterior means. The Wang et al. estimator median MSE is 0.0003, while the posterior means have median MSE 0.0909. The mean MSE, on the other hand, of the two methods are similar. The Wang et al. estimator has mean MSE 0.1796 and posterior means have mean MSE 0.1978.

As previously discussed, the motivation and benefit of developing the hierarchical Bayesian model is to obtain measures of variability for the regression coefficients. The 95% credible intervals for coefficients corresponding to a sample of SNPs, across phenotypes, are shown in



(a) (b)

Figure 7.1: Boxplots of Wang et al. estimator and posterior mean bias with outliers (a) and without outliers (b).



(a) (b)

Figure 7.2: Boxplots of Wang et al. estimator and posterior mean MSE with outliers (a) and without outliers (b).

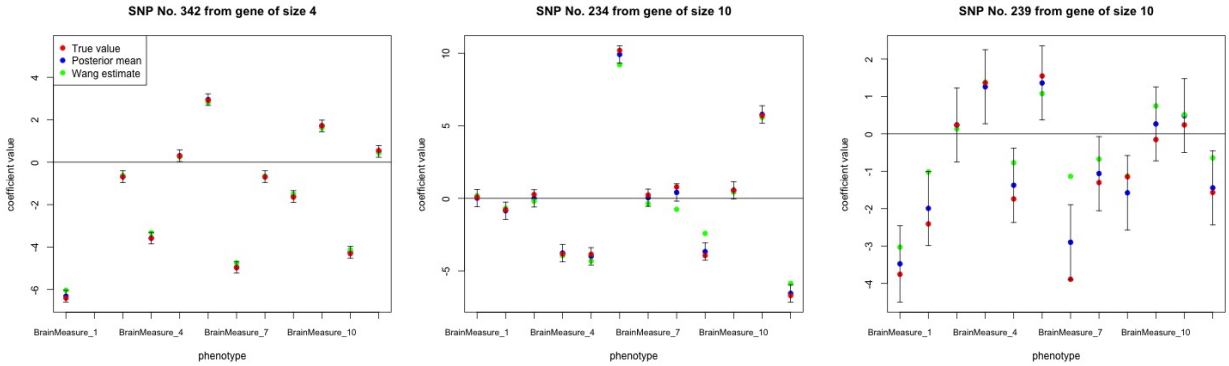


Figure 7.3: Wang et al. estimators (green), posterior means (blue) with 95% credible intervals, and true values (red) for coefficients across phenotypes for 3 SNPs corresponding to rows of $\boldsymbol{W}$ not set to zero.

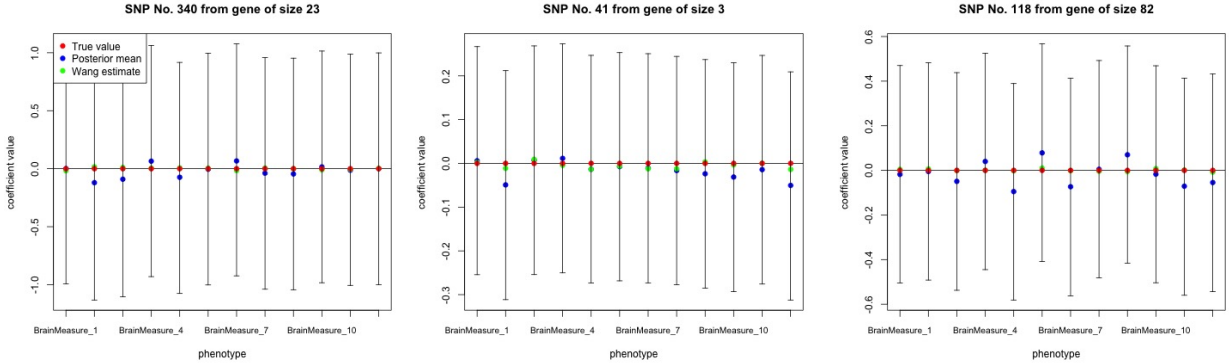


Figure 7.4: Wang et al. estimators (green), posterior means (blue) with 95% credible intervals, and true values (red) for coefficients across phenotypes for 3 SNPs corresponding to rows of $\boldsymbol{W}$ set to zero.

Figures 7.3 and 7.4, the former showing SNPs with true coefficients values not equal to zero and the latter with true coefficients values equal to zero. The performance of the credible intervals is summarised in Figure 7.5. The mean of coverage probabilities from all coefficients is 95.18%. Including only true zero coefficients results in mean coverage probability of 96.54%, while the mean coverage probability of true non-zero coefficients is 83.23%. Figure 7.6 illustrates these same properties, but the outliers are not shown. From this figure, the medians are better depicted. The median of coverage probabilities from all coefficients is 97%,

true zero coefficients median coverage probability is 98%, and median coverage probability of true non-zero coefficients is 90%. In this setting, the Bayesian intervals have reasonably adequate frequentist coverage.

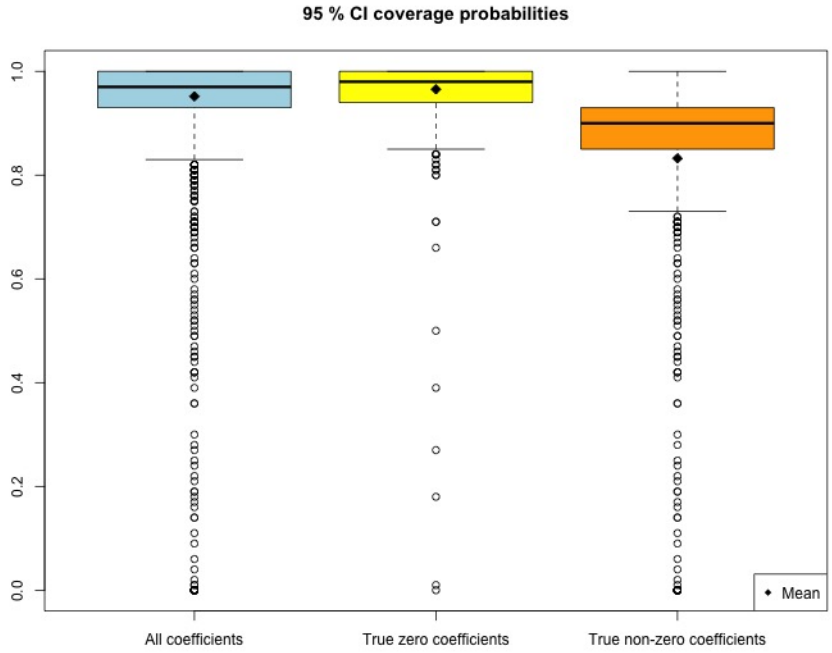


Figure 7.5: Boxplots of 95% credible interval coverage probabilities.

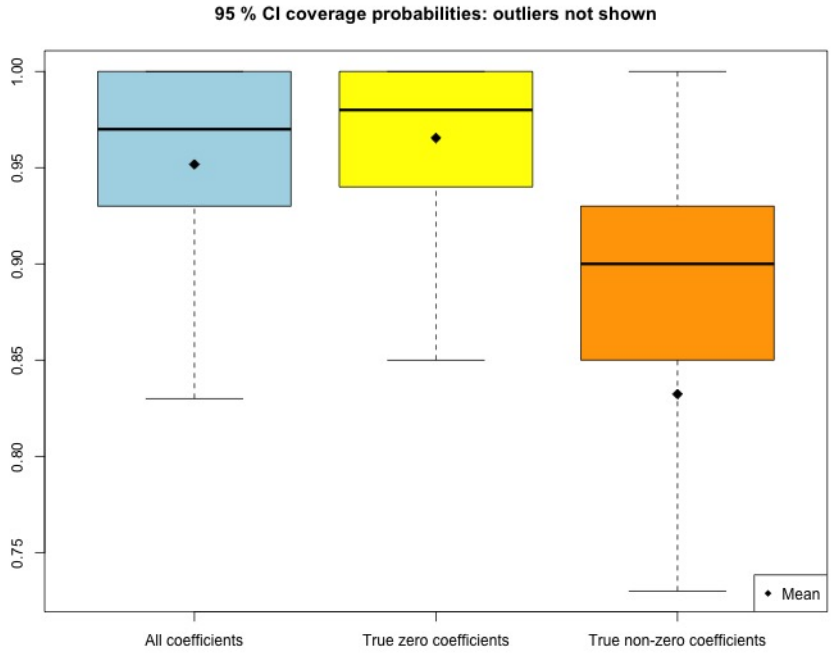


Figure 7.6: Boxplots of 95% credible interval coverage probabilities without outliers.

7.2 Application

The Wang et al. method and our Gibbs-WAIC Bayesian method are applied to the complete dataset outlined in Chapter 2.

Method

The Wang et al. estimator is computed in the same way as in the simulation study.
The Bayesian model is fit with fixed λ_1^2, λ_2^2 values over a two-dimensional grid of $\{10^{-3}, 10^{-2}, \ldots, 10^2, 10^3\}^2$ for a total of 49 MCMC runs, but otherwise includes the same σ^2 prior of $\sigma^2 \sim Inv - Gamma(3, 1)$ and follows the same model fitting procedures as in the simulation study and outlined in Section 7.1.1, where the model with the minimum WAIC is selected.

Results

The Wang et al. method selects tuning parameter values of $\gamma_1 = 10^2$ and $\gamma_2 = 10^2$. The Gibbs-WAIC approach selects $\lambda_1^2 = 10^3$ and $\lambda_2^2 = 10^3$. We note that the selected λ_1^2 and λ_2^2 values are on the boundary of those that were in included for investigation, which may contribute to some of the differences between regression coefficient shrinkage when comparing the Wang et al. and posterior mean estimators.

In regards to Bayesian model SNP selection, there are 5 SNPs that have any 95% credible intervals that do not contain zero, these SNPs are shown in Figure 7.7.

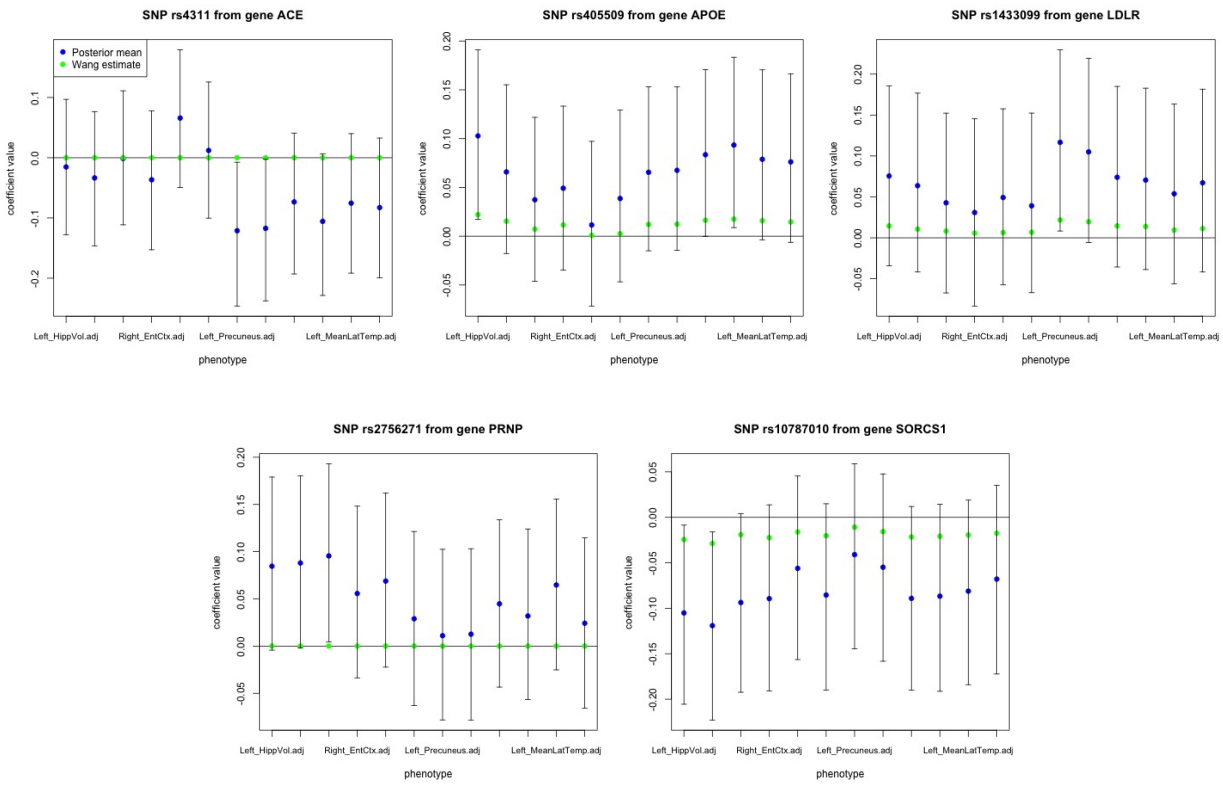


Figure 7.7: Bayesian model selected SNPs with Wang et al. estimates (green), posterior means (blue), and 95% credible intervals.

Wang et al. assign weights to each SNP by summing the absolute values of the estimated

coefficients of a single SNP over all phenotypes. SNPs are then ranked based on their weights. Plots with the top 5 ranked SNPs based on this method are displayed in Figure 7.8. Note that 4 out of 5 of these selected SNPs have 95% credible intervals from our Bayesian-WAIC method that all cover zero. The exception being SNP rs10787010 from gene SORCS1. The results of SNP selection from both methods are summarised in Table 7.1.

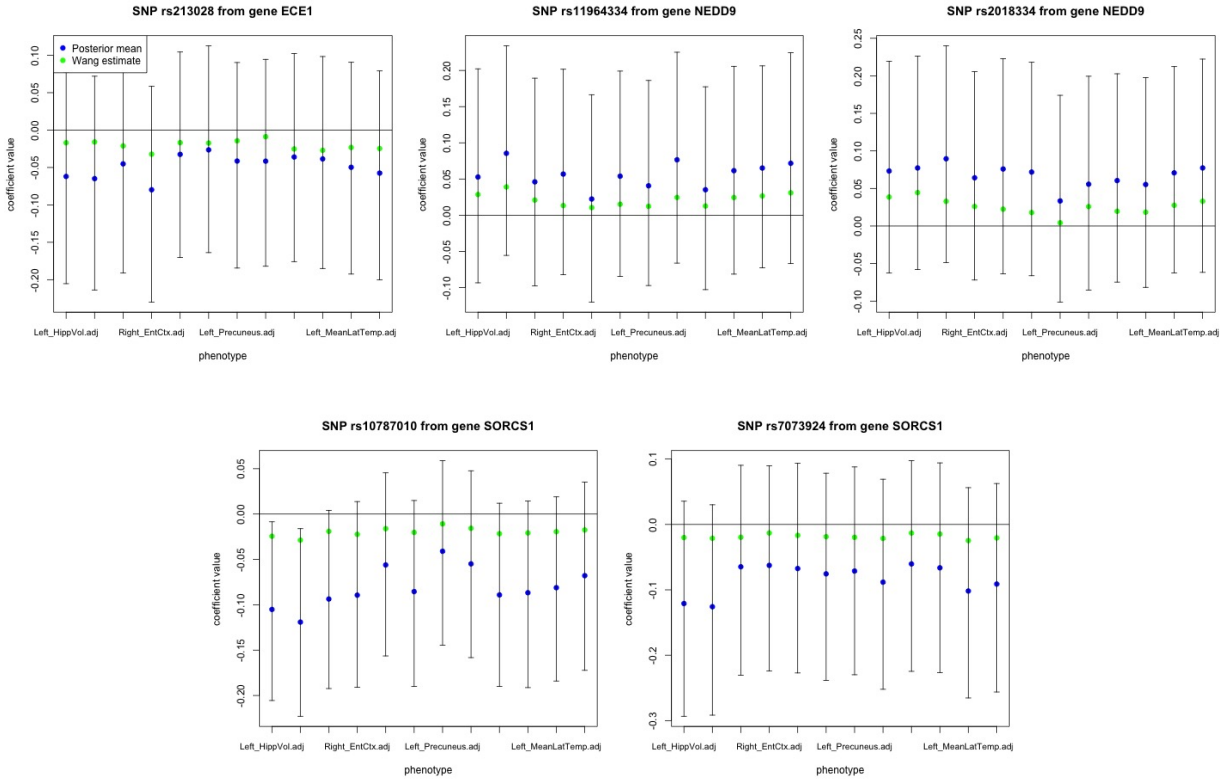


Figure 7.8: Top 5 Wang et al. ranked SNPs with Wang et al. estimates (green), posterior means (blue), and 95% credible intervals.

SNP	Gene	Method of selection
rs4311	ACE	Bayesian
rs405509	APOE	Bayesian
rs1433099	LDLR	Bayesian
rs2756271	PRNP	Bayesian
rs10787010	**SORCS1**	**Both methods**
rs213028	ECE1	Wang et al.
rs11964334	NEDD9	Wang et al.
rs2018334	NEDD9	Wang et al.
rs7073924	SORCS1	Wang et al.

Table 7.1: Bayesian selected SNPs and top 5 Wang et al. ranked SNPs.

Chapter 8

Discussion and future directions

In this work, we have developed a hierarchical Bayesian model so that the mode of the posterior distribution of $\boldsymbol{W}$, the matrix consisting of regression coefficients, is equivalent to the estimator proposed by Wang et al. (2012). This hierarchical model yields full conditionals that take standard forms and Gibbs sampling can be implemented for posterior sampling. The selection of tuning parameters for cases where the number of SNPs is approximately equal to, or outnumbers observations led us to some interesting problems and investigations. After attempting several methods, we conclude that for this model, under settings where the number of SNPs is large relative to the number of observations, choosing tuning parameters based on the marginal likelihood leads to over-shrinking of regression coefficients.

One approach that we did try, but has not yet been mentioned, is to include tuning parameters in the Gibbs sampler but assign very tight priors. Typically the gamma priors used for tuning parameters were quite vague, with variances of at least 1000. When exploring hyperprior options, we did find that assigning very tight priors (eg. variance of less than 10^{-4}) limits the over-shrinking of regression coefficients by forcing λ_1^2 and λ_2^2 estimates to

converge to smaller values than with vague priors. We found, however, through simulation studies that as the number of SNPs included in the model grew larger relative to the number of observations, we had to make the variance of priors smaller, and smaller, to avoid over-shrinking the regression coefficients. We eventually concluded that, for the time being, approximating cross-validation with WAIC was a more reasonable and less arbitrary way to perform model fitting. This is, nonetheless, an area that could be explored in future work. In particular, the work of Castillo, Schmidt-Hieber, and Van DER Vaart (2015) suggests the use of 'spike and slab' priors. Castillo et al. (2015) state that in the case of the Bayesian lasso, the tuning parameter must tend toward infinity in order to shrink regression coefficients to zero. This is certainly a behaviour that we have seen in our model and thus investigating their solution to this problem would be good next step.

After settling on the selection of tuning parameters based on WAIC for our Bayesian model, we evaluate our method with a simulation study and compare performance of posterior means to the Wang et al. estimator. We find similar results across the two methods, and find the additional information provided by a Bayesian framework, the interval estimates, to be quite valuable.

Another possibility for future direction is in the covariance structure of phenotypes. At the first level of our hierarchical model, the imaging phenotypes are assumed to be independent with equal variance. Subsequent work should look at modelling a more realistic covariance structure, including dependence among imaging traits in this first level of the hierarchy.

Appendix A

Appendix

A.1 Notation

d	number of SNPs
c	number of imaging phenotypes
n	number of observations
K	number of genes
π_k	notation for gene k
m_k	number of SNPs in gene k
$\boldsymbol{x}_\ell$	vector of genetic data for subject ℓ
$\boldsymbol{X}$	$d \times n$ matrix of genetic data
$\boldsymbol{y}_\ell$	vector of imaging phenotype data for subject ℓ
$\boldsymbol{Y}$	$c \times n$ matrix of imaging phenotype data
$\boldsymbol{W}$	$d \times c$ matrix of regression coefficients
γ_1, γ_2	tuning parameters of Wang et al. (2012) estimator

$\boldsymbol{W}^{(k)}$	submatrix of $\boldsymbol{W}$ consisting of all rows of $\boldsymbol{W}$ that correspond to the SNPs included in π_k
$\boldsymbol{W}^{(-k)}$	submatrix of $\boldsymbol{W}$ consisting of all the rows of $\boldsymbol{W}$ that correspond to the SNPs not included in π_k
$\boldsymbol{x}_\ell^{(k)}$	vector consisting of all entries of $\boldsymbol{x}_\ell$ that correspond to SNPs included in π_k
$\boldsymbol{x}_\ell^{(-k)}$	vector consisting of all entries of $\boldsymbol{x}_\ell$ that correspond to SNPs not included in π_k
$\boldsymbol{\zeta}^2$	vector $(\zeta_1^2, \ldots, \zeta_{\mathrm{d}}^2)^T$, where ζ_{i}^2 is a parameter associated with the i^{th} SNP
$\boldsymbol{\tau}^2$	vector $(\tau_1^2, \ldots, \tau_{\mathrm{K}}^2)^T$, where τ_k^2 is a parameter associated with the k^{th} gene
$k(i)$	for each $i \in \{1, \ldots, d\}$, $k(i) \in \{1, \ldots, K\}$ denotes the gene associated with the i^{th} SNP
$\boldsymbol{w}^i$	the i^{th} row of $\boldsymbol{W}$

A.2 Background definitions

The definitions of a few genetic terms are provided here. Definitions are taken directly from the following source.

National Institutes of Health (2015). National Human Genome Research Institute. "Talking Glossary of Genetic Terms." Accessed June 2015, from http://www.genome.gov/glossary/.

DNA. "The hereditary material in humans and almost all other organisms is deoxyribonucleic acid (DNA). Nearly every cell in a person's body has the same DNA. The information in DNA is stored as a code made up of four chemical bases: adenine (A), guanine (G), cytosine (C), and thymine (T). Human DNA consists of about 3 billion bases, and more than 99 percent of those bases are the same in all people. The order, or sequence, of these bases determines the information available for building and maintaining an organism."

Single nucleotide polymorphisms. "SNPs are the most common type of genetic variation among people. Each SNP represents a difference in a single DNA building block, called a nucleotide. For example, a SNP may replace the nucleotide cytosine (C) with the nucleotide thymine (T) in a certain stretch of DNA. SNPs occur normally throughout a person's DNA. They occur once in every 300 nucleotides on average, which means there are roughly 10 million SNPs in the human genome. When SNPs occur within a gene or in a regulatory region near a gene, they may play a more direct role in disease by affecting the gene's function. Most SNPs have no effect on health or development. Some of these genetic differences, however, have proven to be very important in the study of human health. Researchers have

found SNPs that may help predict an individual's response to certain drugs, susceptibility to environmental factors such as toxins, and risk of developing particular diseases. SNPs can also be used to track the inheritance of disease genes within families."

Genes. "A gene is the basic physical and functional unit of heredity. Genes, which are made up of DNA, act as instructions to make molecules called proteins. In humans, genes vary in size from a few hundred DNA bases to more than 2 million bases. The Human Genome Project has estimated that humans have between 20,000 and 25,000 genes. Every person has two copies of each gene, one inherited from each parent. Most genes are the same in all people, but a small number of genes (less than 1 percent of the total) are slightly different between people. Alleles are forms of the same gene with small differences in their sequence of DNA bases. These small differences contribute to each person's unique physical features."

Phenotypes. "The observable physical and/or biochemical characteristics of the expression of a gene; the clinical presentation of an individual with a particular genotype. A phenotype is an individual's observable traits, such as height, eye color, and blood type. The genetic contribution to the phenotype is called the genotype. Some traits are largely determined by the genotype, while other traits are largely determined by environmental factors."

References

Bishop, C. M. (2006). *Pattern recognition and machine learning*. springer.

Castillo, I., Schmidt-Hieber, J., & Van DER Vaart, A. (2015). *Bayesian linear regression with sparse priors*. (Submitted to the Annals of Statistics)

Evgeniou, A., & Pontil, M. (2007). Multi-task feature learning. *Advances in neural information processing systems*, *19*, 41–48.

Ge, T., Feng, J., Hibar, D. P., Thompson, P. M., & Nichols, T. E. (2012). Increasing power for voxel-wise genome-wide association studies: the random field theory, least square kernel machines and fast permutation procedures. *Neuroimage*, *63*(2), 858–873.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis*. CRC press.

Gelman, A., Hwang, J., & Vehtari, A. (2014). Understanding predictive information criteria for bayesian models. *Statistics and Computing*, *24*(6), 997–1016.

Kulis, B. (2012). *Conjugate priors.* CSE 788.04:Topics in Machine Learning. http://web.cse.ohio-state.edu/ kulis/teaching/. Accessed March 2015.

Kyung, M., Gill, J., Ghosh, M., & Casella, G. (2010). Penalized regression, standard errors, and bayesian lassos. *Bayesian Analysis*, *5*(2), 369–411.

Levine, R. A., & Casella, G. (2001). Implementations of the monte carlo em algorithm. *Journal of Computational and Graphical Statistics*, *10*(3), 422–439.

McEntyre J, Ostell J, editors. (2015). *The NCBI Handbook [Internet].* Bethesda

(MD): National Center for Biotechnology Information (US); 2002-. Glossary. http://www.ncbi.nlm.nih.gov/books/NBK21106/. Accessed June 2015.

National Institutes of Health. (2015). National Human Genome Research Institute., Talking Glossary of Genetic Terms. http://www.genome.gov/glossary. Accessed June 2015.

National Library of Medicine (US). (2015). *Genetics home reference [internet].* Bethesda (MD): The Library; http://ghr.nlm.nih.gov/. Accessed June 2015.

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

Szefer, E. (2014). *Joint analysis of imaging and genomic data to identify associations related to cognitive impairment.* Unpublished master's thesis, Simon Fraser University, Canada.

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

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., & Knight, K. (2005). Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *67*(1), 91–108.

Wang, H., Nie, F., Huang, H., Kim, S., Nho, K., Risacher, S. L., ... others (2012). Identifying quantitative trait loci via group-sparse multitask regression and feature selection: an imaging genetics study of the adni cohort. *Bioinformatics*, *28*(2), 229–237.

Watanabe, S. (2010). Asymptotic equivalence of bayes cross validation and widely applicable information criterion in singular learning theory. *The Journal of Machine Learning Research*, *11*, 3571–3594.

Yuan, M., & Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *68*(1), 49–67.

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net.

Journal of the Royal Statistical Society: Series B (Statistical Methodology), *67*(2), 301–320.