The LASSO Linear Mixed Model for Mapping Quantitative Trait Loci

Scott Foster

Doctor of Philosophy
May 2006

Supervisors: Arūnas Verbyla and Wayne Pitchford

THE UNIVERSITY OF ADELAIDE
School of Agriculture and Wine
BiometricsSA and Agricultural and Animal Sciences

Contents

1 Introduction **1**

1.1 Overview of Literature 1

1.1.1 QTL Mapping 2

1.1.2 Statistical Methods 3

1.2 Motivating Data 4

1.3 Overview of Thesis 4

2 Review: Genetics and QTL Mapping **7**

2.1 Introduction 7

2.2 Basic (Mendelian) Genetics 7

2.2.1 Constituents of Phenotype 7

2.2.2 Genetic Pathways 8

2.2.3 Genetic Structure 8

2.2.4 Genetic Effect - Terminology 9

2.2.5 Meiosis, Recombination and Genetic Distance 9

2.2.6 F_2 and Back-Cross Populations for Outbreeding Organisms 12

2.3 QTL Mapping In Outbreeding Organisms 13

2.3.1 Missing and Non-Informative Markers 14

2.3.2 Parental Haplotype Reconstruction 14

2.3.3 Single Marker Models 15

2.3.4 Interval Mapping 17

2.4 Multiple QTL methods 23

2.4.1 Multiple Marker Models 23

2.4.2 Composite Interval Mapping 25

2.4.3 Multiple Interval Mapping 26

2.5 Bayesian Methods 27

2.5.1 Single Marker Models 27

2.5.2 Incorporating Multiple Markers 27

2.5.3 Outbred Line Crosses 28

2.5.4 Multiple QTL Models 28

2.5.5 Epistatic QTL Models 29

2.5.6 Performance of Bayesian QTL Methods 29

3 Review: Mixed Models 31

3.1 Introduction 31

3.2 Mixed Models 31

3.3 Restricted Maximum Likelihood 33

3.4 Estimation of Fixed Effects and Variance Parameters 34

3.4.1 Fixed Effects 34

3.4.2 Variance Parameters 35

3.5 Prediction of Random Effects 36

3.5.1 Predictions to Minimise Prediction Error 37

3.5.2 Mixed Model Equations 38

3.5.3 Residuals 40

3.6 Computation for the Standard Mixed Model 40

3.6.1 Working Variates 41

3.6.2 Alternative Forms for Restricted Likelihood and Scores 42

3.6.3 The AI Algorithm 44

3.7 Analytical Approximations to the Marginal Likelihood 44

3.7.1 Laplace's Method 45

3.7.2 Partial Laplace's Method 45

4 Review: Subset Selection and Biased Estimation 49

4.1 Introduction 49

4.2 Why Select Subsets? 49

4.2.1 Prediction 50

4.2.2 Interpretation of Model 52

4.3 Algorithms for Finding Reduced Linear Models 52

4.3.1 Forward Selection 52

4.3.2 Backward Selection 53

4.3.3 Stepwise Selection 54

4.3.4 Best Subsets 55

4.3.5 One True Model? 56

4.4 Ill-Conditioning 56

4.5 Biased Estimation 57

4.5.1 Constrained Likelihood 58

4.5.2 Penalised Likelihood 59

4.5.3 Random Effects Models 60

4.6 Computation for Ridge Regression 61

4.7 Computation for the LASSO 62

4.7.1 First Order Conditions 62

4.7.2 Linearisation and Optimal Descent Directions 63

4.7.3 Interior Point Descent Algorithm . . . 64
4.8 Biased Estimate Standard Errors . . . 66
4.8.1 Ridge Regression . . . 66
4.8.2 The LASSO . . . 67

5 Some Results for the LASSO Random Model 69
5.1 Introduction . . . 69
5.2 Random Effects Model with Single Effect . . . 69
5.2.1 Marginal Distribution of Observations . . . 70
5.2.2 Predictive Distribution . . . 71
5.3 Vector of Effects . . . 81
5.3.1 Marginal Distribution of Observations . . . 81
5.4 Reduction of LASSO Estimates . . . 82
5.5 Summary . . . 84

6 LASSO Computing 85
6.1 Introduction . . . 85
6.2 Interior Point Descent Algorithm for Correlated Data . . . 86
6.2.1 First Order Conditions . . . 86
6.2.2 Linearisation and Optimal Descent Directions . . . 87
6.2.3 Interior Point Descent Algorithm . . . 87
6.2.4 Variable Updating . . . 88
6.3 Estimating Penalised Regression . . . 88
6.3.1 Computation Costs for Penalised Regression . . . 90
6.4 Discussion of Degrees of Freedom for GCV . . . 93
6.4.1 Impact of Estimated Degrees of Freedom on GCV Statistic . . . 94
6.5 Summary . . . 95

7 The LASSO as a Random Effects Model 97
7.1 Introduction . . . 97
7.2 Approximate Likelihood . . . 98
7.3 Derivatives of LASSO Estimates . . . 99
7.4 Score Equations and Estimates . . . 101
7.4.1 Bias in Score Equations and Estimates . . . 103
7.5 Adjusted Score Equations . . . 104
7.5.1 Theoretical Expected Scores . . . 105
7.5.2 Empirical Expected Scores . . . 108
7.6 Estimation Algorithm Overview . . . 109
7.6.1 Number of Bootstrap Samples . . . 110
7.7 Significance Testing of Estimates . . . 111
7.8 Simulation Study . . . 114

7.8.1 Simulation Design . . . 115
7.8.2 Simulation Results . . . 115
7.9 Example: Prostate Data . . . 118
7.10 Summary . . . 121

8 The LASSO Linear Mixed Model 123
8.1 Introduction . . . 123
8.2 The LASSO Mixed Model Equations . . . 124
8.3 Approximate Likelihoods . . . 127
8.3.1 Laplace Approximation . . . 128
8.3.2 Partial Laplace Approximation . . . 129
8.4 Derivatives of LASSO Estimates . . . 132
8.5 Estimation of Dispersion Parameters . . . 133
8.5.1 Unadjusted Score Equations . . . 133
8.5.2 Observed Information . . . 134
8.5.3 Adjusted Score Equations and Estimates . . . 135
8.5.4 Solving Adjusted Score Equations . . . 137
8.5.5 Algorithm Overview . . . 138
8.6 Significance Testing of LASSO Effects . . . 138
8.7 Simulation . . . 139
8.8 Summary . . . 140

9 QTL Detection Using the LLMM 141
9.1 Introduction . . . 141
9.2 QTL Mapping Method . . . 142
9.3 Simulations . . . 143
9.3.1 Heritability 50% . . . 144
9.3.2 Heritability 10% . . . 145
9.3.3 Simulation Summary . . . 145
9.4 Davies' Gene Mapping Data . . . 146
9.5 Summary . . . 149

10 An Alternative LASSO Random Effects Model 151
10.1 Introduction and Motivation . . . 151
10.2 The Alternative Random Effects Model . . . 152
10.3 Approximate Marginal Likelihood . . . 153
10.3.1 Estimating the Random Dispersion Effects . . . 156
10.3.2 An Example of the Approximate Likelihoods' Surfaces . . . 157
10.4 Score Equations and Their Components . . . 158
10.5 Estimation . . . 161
10.6 Performance of Estimation Methods . . . 163

10.6.1 Efficacy 163
10.6.2 Comparison with Bootstrap Adjustment Method 164
10.7 Prediction of LASSO Random Effects 165
10.7.1 Best Linear Predictor 166
10.7.2 Best Predictor 166
10.7.3 Conditional Prediction 167
10.7.4 Example of Different Predictors 170
10.7.5 Performance of Predictors 172
10.8 Inference for Random LASSO Effects 173
10.8.1 Example of Probability Calculation 174
10.9 Re-Analysis of the Prostate Cancer Data 174
10.10Summary 175

11 An Alternative LLMM 179
11.1 Approximate Restricted Likelihood 180
11.2 Score Equations and Their Components 183
11.3 Computation 185
11.3.1 Estimating the Random Dispersion Effects 185
11.3.2 Estimation of Dispersion Parameters 185
11.4 Prediction and Estimation of Effects 186
11.5 Inference for Random Effects 187
11.6 Summary 187

12 Closing Remarks 189
12.1 Overview 189
12.2 Summary of Methodology 189
12.3 Possible Future Research 191
12.3.1 Reducing Computation 191
12.3.2 All-marker Model 191
12.3.3 Distributional Assumptions 192
12.3.4 Multiple QTL Distributions 192
12.3.5 Multiple Traits 193
12.3.6 Outlier Detection 194
12.3.7 Summary of Future Work 195

A Statistical Results 197

B Vector and Matrix Algebra Results 203

C Miscellaneous Results 207

D S-PLUS Functions 209

4.1 Constrained Regression for Dependent Data 209

4.2 Penalised Regression for Dependent Data 213

4.3 LASSO Mixed Model Equations 214

4.4 LASSO Random Effects Model and LLMM 216

4.5 Hypothesis Testing 224

4.6 Alternative LASSO Random Effects Model 226

4.7 Importance Sampling 231

Abstract

This thesis concerns the identification of quantitative trait loci (QTL) for important traits in cattle line crosses. One of these traits is birth weight of calves, which affects both animal production and welfare through correlated effects on parturition and subsequent growth. Birth weight was one of the traits measured in the Davies' Gene Mapping Project. These data form the motivation for the methods presented in this thesis.

Multiple QTL models have been previously proposed and are likely to be superior to single QTL models. The multiple QTL models can be loosely divided into two categories: 1) model building methods that aim to generate good models that contain only a subset of all the potential QTL; and 2) methods that consider all the observed marker explanatory variables. The first set of methods can be misleading if an incorrect model is chosen. The second set of methods does not have this limitation. However, a full fixed effect analysis is generally not possible as the number of marker explanatory variables is typically large with respect to the number of observations. This can be overcome by using constrained estimation methods or by making the marker effects random.

One method of constrained estimation is the least absolute selection and shrinkage operator (LASSO). This method has the appealing ability to produce predictions of effects that are identically zero. The LASSO can also be specified as a random model where the effects follow a double exponential distribution.

In this thesis, the LASSO is investigated from a random effects model perspective. Two methods to approximate the marginal likelihood are presented. The first uses the standard form for the double exponential distribution and requires adjustment of the score equations for unbiased estimation. The second is based on an alternative probability model for the double exponential distribution. It was developed late in the candidature and gives similar dispersion parameter estimates to the first approximation, but does so in a more direct manner.

The alternative LASSO model suggests some novel types of predictors. Methods for a number of different types of predictors are specified and are compared for statistical efficiency.

Initially, inference for the LASSO effects is performed using simulation. Essentially, this treats the random effects as fixed effects and tests the null hypothesis that the effect is zero. In simulation studies, it is shown to be a useful method to identify important effects. However, the effects are random, so such a test is not strictly appropriate. After the specification of the alternative LASSO model, a method for making probability statements about the random effects being above or below zero is developed. This method is based on the predictive distribution of the random effects (posterior in Bayesian terminology).

The random LASSO model is not sufficiently flexible to model most QTL mapping data.

Typically, these data arise from large experiments and require models containing terms for experimental design. For example, the Davies' Gene Mapping experiment requires fixed effects for different sires, a covariate for birthdate within season and random normal effects for management group. To accommodate these sources of variation a mixed model is employed. The marker effects are included into this model as random LASSO effects. Estimation of the dispersion parameters is based on an approximate restricted likelihood (an extension of the first method of estimation for the simple random effects model). Prediction of the random effects is performed using a generalisation of Henderson's mixed model equations.

The performance of the LASSO linear mixed model for QTL identification is assessed via simulation. It performs well against other commonly used methods but it may lack power for lowly heritable traits in small experiments. However, the rate of false positives in such situations is much lower. Also, the LASSO method is more precise in locating the correct marker rather than a marker in its vicinity. Analysis of the Davies' Gene Mapping Data using the methods described in this thesis identified five non-zero marker-within-sire effects (there were 570 such effects). This analysis clearly shows that most of the genome does not affect the trait of interest.

The simulation results and the analysis of the Davies' Gene Mapping Project Data show that the LASSO linear mixed model is a competitive method for QTL identification. It provides a flexible method to model the genetic and experimental effects simultaneously.

Declaration

This work contains no material which has been accepted for the award of any other degree or diploma in any university or other tertiary institution and, to the best of my knowledge and belief, contains no material previously published or written by another person, except where due reference has been made.

I give consent to this copy of my thesis, when deposited in the University Library, being available in all forms of media, now or hereafter known.

SIGNED: .... **DATE:** ....

Acknowledgements

Firstly and most importantly, I would like to thank my supervisors Ari and Wayne. Not only have they been very generous with their advice, knowledge and patience, but also with their friendship. I have learnt both technical and personal skills from them that I will endeavour to carry through life. It has been a real pleasure.

The seminal idea for this thesis was passed on from Robin Thompson. I am indebted as it has proved interesting, challenging and rewarding.

The data set from the Davies' Gene Mapping Project was generously made available by Wayne and Cindy Bottema. It represents many years of hard work by their team and I appreciate the chance to use it.

I would like to thank the staff and students at BiometricsSA and in the Discipline of Agricultural and Animal Sciences. They have provided support, technical advice and stimulating discussions. Julian Taylor has been especially helpful (and has been very gracious about giving time for 'dumb' computing questions).

My scholarship was provided in part by the University of Adelaide and in part by the Co-operative Research Centre for Beef and Cattle Quality. I am grateful for their support, without which I would not have even started.

My partner Zoë deserves special thanks. Our time in Adelaide has been truly happy.

BIOMETRICS SA

CRC BEEF QUALITY

Chapter 1

Introduction

A cattle herd improvement strategy is an integral component in maximising profitability of any beef production system. Traditionally, herd improvement has been implemented by measuring every animal for various traits of economic importance and choosing those animals that exhibit favourable phenotypes for breeding purposes. The favourable genes are passed on to the next generation of animals within the herd. However, this strategy is not practical for some traits of economic importance. For example: feed efficiency, the cost of measurement outweighs the potential increase in profitability; and calving ease, the trait is measured after a potentially fatal birth event. An alternative strategy that may work for these traits is to identify genes or proxies for those genes that control these traits and then identify animals with favourable alleles. Thus, herd improvement can occur without direct measurement of the traits.

Incorporating the genetic information into the breeding strategy could take the form of marker assisted selection (MAS) where a molecular marker is identified that is genetically linked to a gene or genes that affect the trait of interest. Such markers are sometimes called quantitative trait loci (QTL singular and plural). The QTL are generally not the causative gene themselves. Hence, care must be taken when using these markers as the linkage phase of the QTL and gene may differ between populations.

QTL can also be used to assist gene discovery. They identify regions of the genome that may contain causative genes. This localisation narrows down the set of candidate genes drastically. Once a causative gene is known it can be used in breeding schemes without needing to know the linkage phase. Also, knowledge of the causative gene enhances understanding of the quantitative trait through its affect on the biochemistry of the animal.

1.1 Overview of Literature

Here, a general overview of the literature is given to motivate the methods developed in this thesis. A more detailed review is given in subsequent chapters. The overview consists of two parts: QTL mapping; and statistical methods relevant to this thesis.

1.1.1 *QTL Mapping*

The basic idea behind QTL mapping in diploid species is to statistically associate a trait of interest to a region of the genome, measured by (molecular) markers. The idea was originally proposed by Sax (1923), for associating quantitative traits to dichotomous measurements on individuals. In outbreeding diploid species, such as cattle, QTL mapping is based on studying the change in the trait means of the two inheritance classes for the different alleles of a locus (Beckmann & Soller, 1988).

The popularity of QTL mapping has escalated with the ease of scoring molecular markers. In essence, molecular markers measure a dichotomous variable on an individual. The appeal of molecular markers is that it is possible to score many different markers on each individual, quickly and cost effectively. This increases the amount of genetic information, enabling more precise QTL mapping to occur.

A simple statistical method that was popular in early QTL mapping was examining the differences in trait values between a single marker's inherited allele types. Each marker is fitted in a separate model and the QTL was deemed to be associated with the marker that produced the best model. Typically the QTL effect is assumed to be equal to the marker effect, even though the marker effect will be attenuated the further the marker is from the QTL (Geldermann, 1975). To overcome this, Weller (1986) proposed a mixture model based on a single marker. This model acknowledges that there is a possibility of cross-overs between marker and gene and hence the marker classes may not be the gene's classes. Lander & Botstein (1989) extended this model by incorporating knowledge about how the markers were related to each other (using a linkage map). Their method scans the entire genome using consecutive pairs markers as the basis for identifying the inheritance classes at a given genomic position. This method is called interval mapping (IM) and the mixture model at each putative location is estimated using maximum likelihood. This estimation procedure can be computationally expensive and some less demanding methods have been proposed. The most popular of these is the regression interval mapping method (Haley & Knott, 1992 and Martinez & Curnow, 1992). Whittaker et al. (1996) showed that computation can be further reduced by noting that the QTL additive effect is a function of the flanking markers' effects.

The marker-by-marker and IM methods consider a series of models that contain a single QTL. The single QTL assumption is unlikely to be true as quantitative traits are generally thought to have a complex genetic basis (e.g. Hartl & Jones, 1998 and Kearsey & Pooni, 1996). The single QTL models can be extended to allow for multiple QTL using one of two common themes. The first is to use models that contain multiple markers (or intervals), selected from the full set. The selection can be based on marker explanatory variables only (Broman & Speed, 2002), on a combination of intervals and markers (Jansen, 1993a; Zeng, 1993; Jansen, 1994; Jansen & Stam, 1994; and Zeng, 1994), or on intervals only (Haley & Knott, 1992; Whittaker et al., 1996; Kao et al., 1999; and Zeng et al., 1999). The second theme circumvents the issue of selection and includes all the marker effects as random effects, equivalent in many cases to constrained or penalised estimation (e.g. Lindley & Smith, 1972). Most of these methods (Whittaker et al., 2000; Ball, 2001; Gianola et al., 2003; and Xu, 2003) use marker explanatory variables and not intervals. The method of Xu (2003) has been extended to use intervals (Wang et al., 2005) but this is unlikely to increase precision if the markers are sufficiently dense (Broman & Speed, 2002).

Whittaker et al. (2000) and Gianola et al. (2003) assume that the distribution of the marker effects is normal. Xu (2003) also assumes normality but allows a separate variance for each marker effect; a function of the marker's effect. This allows for differential shrinkage between the markers. Some markers are estimated to be approximately zero as they have very small variance. Thus, the method of Xu (2003) essentially performs model selection using an all marker model.

It is unclear if the normality assumption is appropriate for the marker effects. Empirical evidence suggests that QTL effects have kurtosis similar to a double exponential variable (Chamberlin et al., 2005), higher than a normal variable. The double exponential distribution may provide a better model for the QTL effects as the first four moments match the unknown distribution.

1.1.2 *Statistical Methods*

Recently, there have been a number of estimation methods proposed that perform selection. Often, these are posed as constrained regression problems where the constraint forms a solution space whose shape increases the chance of obtaining identically zero estimates. These methods are: the non-negative garrote (Breiman, 1995); the least absolute selection and shrinkage operator (LASSO; Tibshirani, 1996); the bridge (Frank & Friedman, 1993 and Fu, 1998); and the elastic net (Zou & Hastie, 2005). They are similar in construction to the popular ridge regression (Hoerl & Kennard, 1970a and Hoerl & Kennard, 1970b), although ridge regression does not produce identically zero estimates. Ridge regression can be formulated as a random effects model where the effects and residuals are normally distributed (Lindley & Smith, 1972). The LASSO has a similar formulation; the effects are double exponentially distributed (Tibshirani, 1996).

As noted by Yuan & Lin (2005) the double exponential distribution has an alternative formulation. If β is a centred double exponential distribution with variance $2\phi^2$ then it can be expressed as the margin of joint distribution $f(\beta|\delta)f(\delta)$ where $f(\beta|\delta)$ is a centred normal distribution with variance δ and $f(\delta)$ is an exponential distribution (Andrews & Mallows, 1974 and Kotz et al., 2001). This formulation suggests that the LASSO induces different variances for the different random effects (when considered as a normal model).

The LASSO has generated a considerable amount of interest amongst researchers. This is perhaps due to its conceptual simplicity. There have been a number of algorithmic methods proposed to estimate the effects given a constraint or penalty parameter (Tibshirani, 1996; Fu, 1998; Osborne et al., 2000; Öjelund et al., 2001; Miller, 2002; and Efron et al., 2004). However, it is unclear how to best choose the constraint or penalty parameter. Suggestions include cross-validation (Tibshirani, 1996), generalised cross validation (Tibshirani, 1996 and Fu, 1998) and the use of the C_p statistic (Efron et al., 2004). All these methods require an estimate of the degrees of freedom which varies from one author to another.

1.2 Motivating Data

The Davies' Mapping Project was established with the goal of discovering QTL and causative genes for a variety of traits in beef cattle. The data from the main experiment of this project form the motivation for the methods presented in this thesis. Particular attention is paid to the birth weight trait which affects both animal production and welfare through correlated effects on parturition and subsequent growth rates. Birth weight was measured within the first day of an animal's life. For a more detailed description of the Davies' Gene Mapping Experiment see Afolayan et al. (2002a) and Afolayan et al. (2002b).

The genetic design of the Davies' Experiment was a double back-cross using three F_1 sires, produced from Limousin cattle (a beef breed) and Jersey cattle (a dairy breed). Each sire was used to create a large half-sib family, whose dams were either Limousin or Jersey. In total there were 356 progeny, approximately evenly spread between the sires. There were 287 molecular markers scored at 264 unique genomic positions spanning the 29 cattle autosomes. The markers were mostly microsatellites, but there were also 36 single nucleotide polymorphisms (SNPs) located in 20 genes. Each sire was heterozygous for only a subset of the markers; sire 361 had 188, sire 368 had 184 and sire 398 had 198 heterozygous markers. It is expected that birth weight will depend on sire, breed of dam (BOD) as well as QTL.

Logistics (e.g. semen availability, yard size and labour requirements) meant that a non-genetic experimental design was not only desirable but necessary. In particular, the progeny were produced over three years and the steers and heifers were treated as separate management groups. Each year-by-sex combination is considered to be a level of the Cohort factor. Further, it is expected that birth weight will depend on the date within season that the animal was born. Thus, the Cohort factor and the date of birth covariate should also be accounted for when identifying QTL as these may explain variation in birth weight.

1.3 Overview of Thesis

The aim of this thesis is to assess the use of LASSO methodology for QTL mapping. The methods are developed with the Davies' Gene Mapping data in mind. The genetic effects should be modelled simultaneously with the experimental effects. This is performed by incorporating LASSO random effects into a standard mixed model, which contains fixed and random normal effects. This new model is named the LASSO linear mixed model (LLMM) throughout this thesis.

Reviews of genetics and QTL mapping (Chapter 2), mixed models (Chapter 3) and subset selection and constrained estimation methods (Chapter 4) are provided to motivate the methods in this thesis. They also provide a useful background to the details of the published techniques.

Viewing the LASSO effects as random suggests that the level of shrinkage can be chosen by the ratio of the residual variance and the dispersion parameter for the LASSO effects (Chapter 4). Hence, estimating the dispersion parameters may be a potential method to estimate the level of shrinkage. To accomplish this estimation, the likelihood based on the marginal distribution of the observations is used. Exact analytical forms for a LASSO random effects model (containing only LASSO effects) are presented in Chapter 5. The analytical forms for the marginal distribution are difficult to evaluate and to optimise. Hence, approximations are sought.

Two approximations based on Laplace's method are presented. The first approximation (Chapter 7) produces score equations that are biased. These are adjusted using a computer intensive bootstrapping method (McCullagh & Tibshirani, 1990). The second approximation (Chapter 10) uses the hierarchical form for the double exponential distribution and does not require adjustment to the score equations.

Both these approximations are extended for the LLMM using a modification of Laplace's method (Taylor & Verbyla, 2006) that allows for estimation of fixed effects. These are approximate restricted likelihoods. The first approximation is presented in Chapter 8. Paralleling its random effects model counterpart in Chapter 7, it produces score equations that are biased and are adjusted in a similar manner. The second approximation using the hierarchical formulation of the double exponential distribution is presented in Chapter 11. The second approximation has not been implemented.

After estimation of the dispersion parameters in the random LASSO model the analyst will typically want to predict the random effects. The standard LASSO method is to predict using the mode of the predictive distribution (Tibshirani, 1996). However, this is not the only possibility. For example, the expected value of the predictive distribution should provide predictions with lower mean squared error (e.g. Searle et al., 1992). These issues are investigated for the LASSO random effects model in a theoretical manner in Chapter 5 and empirically in Chapter 10.

The direct extension to standard LASSO prediction in the LLMM is to find the mode of the joint predictive distribution. This is achieved using an extension of Henderson's mixed model equations (Chapter 8). Solving for the LASSO effects can be posed as a penalised regression problem where the residuals have a known correlation structure. This is performed using an extension of the interior point descent algorithm (Osborne et al., 2000), described in Chapter 6.

Two methods of inference for the random LASSO effects are presented. The first treats the LASSO predictions as though they were fixed effects (Chapter 7). An empirical distribution is created assuming that all LASSO effects are zero. The observed predictions are then compared against this distribution. This method of inference is used in the demonstrated QTL mapping methodology and appears to work well. However, it is not strictly appropriate for random effects. A more appropriate method is to calculate the conditional probability of the effect, given the data, being greater than (less than) zero. A method to perform these calculations is given in Chapter 10.

Details for using the LLMM for QTL mapping are presented in Chapter 9. The hierarchical form of the LASSO model suggests some similarities with Xu (2003). In particular each marker effect is a normal random variable with its own variance. The LLMM's performance is assessed via simulation and compared against the recommended method from Broman & Speed (2002).

Chapter 2

Review: Genetics and QTL Mapping

2.1 Introduction

In this Chapter, a review of statistical methods for identifying QTL is presented. However, before any statistical models are proposed an understanding of the elementary genetical processes underlying QTL detection is crucial. These processes are briefly reviewed in the first two sections. The remainder of the Chapter contains a review of published methods used to locate QTL.

2.2 Basic (Mendelian) Genetics

2.2.1 *Constituents of Phenotype*

The phenotype of an animal is its overall physical characteristic. Often individual aspects of an animal's phenotype will be measured. Such measurements are called traits. The words phenotype and traits are often confused. A phenotype is an amalgamation of an infinite number of traits. However, common usage permits phenotype to be used as a synonym for trait.

The phenotype of an animal has constituent parts; genetic, environmental and their interaction. The environmental part is governed by how the physical environment affects the animal. An example is two genetically identical animals, one with poor feed and the other with good feed. The animal on poor feed will have less condition than the animal on good feed. The genetic part of the phenotype is governed by how the animal acts as a biological system. In an equal environment some animals will perform better than others. Realistically, there is a genetic and environmental interaction. The way the animal functions is dependent on the environment that it is placed in. In animal studies this interaction is seldom considered as this would require genetic replication, a practically impossible requirement for outbreeding organisms.

2.2.2 *Genetic Pathways*

The biological function of an animal is governed by how many and what type of proteins it produces. The amount and type of proteins a cell produces is governed by which genes an animal carries. Most types of cells contain complete copies of all of the genes carried by an animal. Each gene is a sequence of deoxyribonucleic acid (DNA), which is embedded in longer sequences called a chromosome. There are many genes located on each chromosome.

Genes form proteins by a process termed the "central dogma", comprising two main processes; the first is transcription and the second is translation. Transcription is the process which takes the gene and generates ribonucleic acid (RNA). The RNA in turn, through translation, is converted into protein(s). For a more thorough description of the central dogma see Hartl & Jones (1998) and Hartwell et al. (2000).

Some genetic effects on the phenotype are best described through the central dogma. Dominance occurs when a single copy of a gene does not produce half the effect in the phenotype as two copies of the gene. This occurs when the amount or type of protein is altered by having a single copy of the gene. Pleiotropy occurs when a protein affects more than one trait measured on the phenotype. Epistasis occurs when the individual genes' effects do not explain the total genetic portion of the phenotype. That is, the effect of one gene is dependent on the genotype at another gene or other genes. There are two causes for epistasis: a gene's protein may not be produced unless another gene's protein is present, or the proteins themselves interact.

2.2.3 *Genetic Structure*

A chromosome is a single strand of DNA that contains genes. Every complex organism has more than one chromosome. For example, humans have 23 chromosomes, fruit fly 4, wheat 7 and cattle 29 (plus one sex chromosome). Humans, fruit fly, cattle and many other species have two copies of each chromosome. Such organisms are called diploid and the two copies of each chromosome are called homologous pairs. In diploid organisms one of the chromosomes is inherited from the father and one from the mother. Some organisms have only one copy of each chromosome (haploid organisms) and some, for example wheat, have more than two copies (polyploid organisms).

A position on a chromosome is called a locus (loci plural) which may or may not be a position in a gene. Each locus represents a position on each of the homologous chromosomes. In diploid organisms, the genetic material at a locus has two components, called alleles, one on each chromosome.

If a locus has the same allele on each of its chromosomes then the locus is said to be homozygous. If the alleles are different at a locus then it is said to be heterozygous at that locus. An animal that is homozygous at each and every locus along a chromosome is said to be inbred for that chromosome. It is inbred if it is inbred for every chromosome. If an animal is not inbred, it is outbred. In reality no animal species is completely inbred, even laboratory animals bred heavily to increase homozygosity still contain heterozygous loci (measured or otherwise). However, theoretical reasoning using an inbred organism model is often useful.

It can be important to know if the alleles of two distinct loci on the same chromosome lie on the same member of a homologous pair. Such information is called the phase of the alleles. If two alleles lie on the same chromosome then they are said to be 'in-phase' (coupling phase). If they are on different chromosomes then they are 'out-of-phase' (repulsive phase).

There are chemicals, called molecular markers, designed to bind to particular loci and whose expression can be measured. If a change in the alleles at that locus is made then the expression of the chemical is also changed, enabling the measurement of which alleles an individual is carrying to be performed.

2.2.4 *Genetic Effect - Terminology*

Consider a gene A that affects a single quantitative trait and is the only gene doing so. Also, suppose that the gene has only two alleles A_1 and A_2. If the trait value is the same for individuals homozygous for A_1 and heterozygotes, then the allele A_1 is termed dominant and A_2 is recessive for the trait. If the trait value for the heterozygote is exactly half way between the trait values for both homozygotes then neither of the alleles is dominant. If the trait value for the heterozygote is in between but closer to the trait value for allele A_1, then A_1 is partially dominant over A_2. The trait value for the heterozygote can also lie outside the interval bounded by the trait values for the homozygotes, this is called over-dominance.

An interesting class of loci are those whose heterozygotes express both the values of the homozygotes. These are called co-dominant loci. Some types of molecular markers (including the ones considered in later chapters) often fall into this category - heterozygotes can be distinguished from homozygotes.

For quantitative traits governed by a single gene, a number of useful summary quantitative values can be defined. These are defined on the means of the different allele combinations for the trait. The mid-parent value is the mean of both homozygote groups, and if there is no dominance then this is also the trait value of the heterozygotes. If there is dominance then it is defined as the distance from the mid-parent to the heterozygote's trait value. The additive effect is the difference between each homozygote group's trait value and the mid-parent, hence the difference between the homozygotes trait values is twice this value. Thus quantitative traits are conceptually handled in a similar manner to categorical traits. For a more in depth discussion of genetic effects for quantitative traits see Kearsey & Pooni (1996, Chapter 1).

2.2.5 *Meiosis, Recombination and Genetic Distance*

The formation of sex cells, called gametes, is completed in the meiosis process. This process creates new chromosomes which contain segments of the original homologous pair. Each created chromosome is unique and has no homologous counterpart. An abridged description of meiosis is now given, a more complete description is given in Hartl & Jones (1998).

- Each chromosome in a homologous pair doubles to form sister chromatids. The sister chromatids are joined.
- The homologous pairs align to form a bivalent pair, a series of four chromatids - two pairs of sister chromatids.
- During the joining of the pairs of sister chromatids the non-sister chromatids can touch, forming a chiasma. At these points both the non-sister chromatids break and recombine with the remaining portion of the chromosome, in effect swapping an equal-in-length portion of the chromosome. The positions where the non-sister chromatids swap DNA is called a 'recombination event' or simply 'recombination'.
- The recombined chromatids separate and form haploid sex cells; the gametes.

It is the gametes, recombined chromosomes, that are inherited from a parent to its offspring rather than a complete copy of one of the parent's chromosomes. The offspring also receives a set of chromosomes from the other parent, which have also potentially undergone recombination in meiosis. This is the method that a parent passes on recombined copies of its genetic material.

Two alleles of loci on different chromosomes will be inherited together with stochastic independence, as the formation of gametes is independent for different chromosomes. If the two loci are on the same chromosome and are in coupling phase then the alleles will be inherited together unless there is a recombination event between them. The probability of such a recombination event diminishes with the distance between the loci. As these loci will have a higher probability of being inherited together, they are said to be linked.

It is commonly assumed that the probability of occurrence of a recombination event at a particular locus is independent of the presence of another recombination event elsewhere on the chromosome. If this is the case then the recombinations are said to occur with no interference. The opposite assumption is that the presence of a recombination event somewhere on the chromosome prevents another from occurring, called complete interference. Haldane (1919) showed that neither assumption is valid, at least for the model organism, the fruit fly.

During many meioses, the number of recombination events between two loci can be recorded. The fraction of the recombinations to the number of meioses is a useful measure to assess if the loci are linked and if they are then how tightly. This fraction is called the 'recombination fraction'. If the loci are unlinked then the expected recombination fraction will be 0.5 and if the loci are closely linked then it will be a small positive fraction.

Two recombination events along a chromatid in the interval flanked by two observed loci are possible. The loci alleles will still be in coupling phase even though recombinations have occurred between them. This implies that the recombination fraction does not capture all the recombination events that take place. In fact the recombination fraction is an under estimate.

The recombination fraction estimates are non-additive over neighbouring intervals if there is not complete interference. To demonstrate, suppose that A, B and C are three loci, all on the same chromosome and in that order. Let r_{ab} be the recombination fraction between A and B and likewise for r_{bc} and r_{ac}. Then for situations without complete interference $r_{ac} \neq r_{ab} + r_{bc}$. This inequality arises because there may be chromatids that recombine within the first interval and within the second.

To accommodate the double recombinations and the unknown level of interference the following formula can be used

$$r_{ac} = r_{ab} + r_{bc} - \delta r_{ab} r_{bc}$$

where $0 \leq \delta \leq 2$ represents interference. A small value of δ represents complete interference and a large one little interference.

In practise, this model is not used, as the choice of δ is not obvious and the recombination fractions are not linear, making expressions for multiple intervals unwieldy. Commonly, the recombination fractions are converted into genetic distances, with a linear scale. The function to convert recombinations to genetic distances is called a mapping function.

The most popular mapping function is Haldane's mapping function (Haldane, 1919). This function assumes that there is no interference along the chromosome. A summary of the derivation is now given.

Let the loci A and B define a segment of a chromosome such that the genetic distance between the loci is $d(A, B) = d_g$. This segment can be divided up into N small segments of equal length. As $N \to \infty$ the probability of getting exactly t recombination events within the interval flanked by A and B follows a Poisson distribution. This is justified by observing that the outcome is a count given an infinite number of observations. The Poisson probability distribution is

$$\mathrm{P}(t) = \frac{d_g^t e^{-d_g}}{t!}.$$

Double recombinations or any even t recombination events are not observed, so only the odd t are summed over to obtain the probability of observing a recombination event between the loci. This sum gives

$$\begin{aligned} r_{ab} &= \mathrm{P}(1) + \mathrm{P}(3) + \mathrm{P}(5) + \ldots \\ &= e^{-d_g} \left(\frac{d_g^1}{1!} + \frac{d_g^3}{3!} + \frac{d_g^5}{5!} + \ldots \right) \\ &= e^{-d_g} \sinh(d_g) \\ &= \frac{1 - e^{-2d_g}}{2}, \end{aligned}$$

and

$$d_g = -\frac{\log(1 - 2r_{ab})}{2}.$$

The assumption of no interference is incorrect (Haldane, 1919). There are other functions, such as Kosambi's that take into account partial interference (Kosambi, 1944), but these are used less frequently. The reason for this is unclear, it may be due to increased complexity for relatively little inferential gain.

2.2.6 F_2 and Back-Cross Populations for Outbreeding Organisms

It is important to consider some of the common experimental structures used for genetical experimentation. The most common populations are the F_2 cross and the (first) back-cross, with back-crossing performed to potentially both founding parents.

Both of these structures start with identifying two original parental lines that are divergent for some trait of interest; call these P_l and P_h for the low and high lines for the trait respectively. For ease of exposition, suppose that both the divergent parental populations are inbred, that is they are all homozygous for all loci and within each population the animals are all genetically identical. At an arbitrary locus M the allele type of the low line is m_l and the allele type of the high line is m_h. The parental lines are crossed to produce a population of F_1 individuals. The F_1 individuals will be heterozygous and genetically identical at all loci.

The F_1 individuals can then be mated to other F_1 individuals producing the F_2 population. This population will contain individuals that are genetically segregating, that is they have different chromosomal construction due to meiosis. This arises because both F_1 mother and father can pass on either allele. For the arbitrary locus M, the allele types of the F_2 population are $\mathrm{m}_l\mathrm{m}_l$, $\mathrm{m}_l\mathrm{m}_h$, $\mathrm{m}_h\mathrm{m}_l$ and $\mathrm{m}_h\mathrm{m}_h$. Each allele type occurs with frequency of one quarter, implying that one quarter of the animals will be homozygous for the low allele, one quarter homozygous for the high allele and one half will be heterozygous.

Instead of mating the F_1 individuals with themselves, they can be mated to either (or both) of the parental lines. These crosses are called back-crosses. Consider the cross to the low parental line, BC_l. All the BC_l individuals will inherit a m_l allele from their P_l parent and either a m_1 or a m_h allele from their F_l parent. The population structure of the BC_l population is one half homozygotes for the m_l allele and one half heterozygotes. The other back-cross, between the high parental line and the F_1 population, BC_h is likewise found to be one half homozygotes for the m_h allele and one half heterozygotes. A double back-cross population is formed by making back-crosses to both of the parental lines.

The double back-cross will have the same genetical composition as an F_2 cross. However, there is a distinct and important difference. The populations are parented by different genetic types. In the F_2 population both parents are heterozygotes and in the double back-cross the parents come from all three genetical types. If there is any difference in the quality of parenting between the genotypes, for example milk quality, then this may affect the trait values of the offspring. Any dominance measure calculated from a double back-cross population may be estimating parenting skill of the different genotypes.

All the information given on population structures up to this point has assumed inbreeding. This assumption makes the process easier to understand. For outbreeding organisms such as cattle, the original individuals in the divergent lines are not necessarily homozygous, nor are the animals within each line genetically identical to each other. This means that two members of the parental lines (one from each divergent line) may have in common zero, one or two alleles at a given locus. Further, the set of alleles in the high parental line may not be disjoint to the set of alleles in the low parental line. Hence the F_1 individuals at a particular locus may be heterozygous or homozygous for any number of alleles. If these individuals are then used to form an F_2 or back-cross population the progeny may again take any genetical form. However, it is expected (and hoped) that two divergent parental lines will differ for the genes of interest and hence, for at least this locus, the inbreeding model holds.

Animals in a mapping population should ideally only differ genetically from each other due to recombination in meiosis. This implies that the best population to use is one where the meioses from both parents can be examined. This can only be done with full-sib families, where progeny share the same mother and father. Unfortunately in species with low fecundity, such as cattle, such full-sib populations are time consuming and/or expensive to produce. Cattle typically only have one offspring per year and hence full-sib families will span multiple years and the population may possess genotype-by-year interaction. Instead half-sib populations are commonly used where a single parent is mated to many others. The genetic effects are then tracked for the common parents and the genetic effects of the non-common parents are considered as noise.

2.3 QTL Mapping In Outbreeding Organisms

The objective of many genetic experiments is to try and relate a quantitative trait to a locus or loci controlling it. Typically this is done by producing F_2 or back-cross populations from divergent lines that may or may not be inbreeding. The individuals in the derived population have two types of measurements taken on them, a particular trait of the phenotype of interest and a number of molecular markers. These markers form proxies for regions of the genome which are used to assess the associations with the trait.

The idea behind mapping QTL in outbreeding organisms is a slight generalisation of that for inbreeding organisms. Association of genomic regions, measured by markers, and the trait is assessed by considering which alleles, at which loci, carried by the parents cause a change in the trait value within the progeny population (Beckmann & Soller, 1988). Consider an animal that at locus A carries a high and low allele a_h and a_l respectively. All the progeny from this animal that inherit the high allele a_h will, on average, have a higher trait value than those progeny inheriting a_l. For inbreed crosses the allele types imply inheritance type.

The statistical methods presented later in this chapter will assume that all genetic effects of interest can be estimated for the given design. Also, it will be assumed that there are only two alleles. In many outbred populations there will be more than two, however within a half-sib family of diploid organisms there will only be two alleles for which inheritance can be inferred.

2.3.1 *Missing and Non-Informative Markers*

The observed molecular marker scores are observed random variables. These data are subject to measurement error and missing values. Further, in line crosses of outbred organisms it is often impossible to definitively assign which of the progeny's alleles were inherited from which parent (Beckmann & Soller, 1988 and Haley et al., 1994). Both missing marker and non-informative markers can generally be treated in the same manner.

The missing or non-informative markers can be replaced by the expected probability of inheriting one of the parental alleles based on the neighbouring markers and the parental linkage phase (Martinez & Curnow, 1994). Alternatively, when assessing a particular locus for different animals, potentially different markers can be used depending on information content of the animals' nearest markers (Haley et al., 1994). These two approaches can be shown to be equivalent (Paul Eckermann, 2003 *personal communication*).

2.3.2 *Parental Haplotype Reconstruction*

The construction of the two parental haplotypes is crucial for calculating which marker alleles each individual progeny inherited from their parents. Construction for inbred lines is trivial as each parental line has its own alleles which are unique within the experiment. In outbreeding organisms the construction of haplotypes considers pairs of consecutive markers on the linkage map in turn.

Assume that the linkage map is known, either by a consensus map or by construction. Consider the 2 parental haplotypes H_1 and H_2, a population of its progeny Ω and two consecutive markers M_l and M_r with alleles m_{l1}, m_{l2}, m_{r1} and m_{r2} respectively. Assign m_{l1} to lie on parental haplotype H_1. Haplotype reconstruction for these two markers reduces to deciding if the marker allele m_{r1} at marker M_r is in coupling or repulsive phase with m_{l1} and hence if it lies on parental haplotype H_1. Using Bayes' law, the probability of the parental haplotypes are

$$\begin{aligned} \mathrm{P}(m_{l1}//m_{r1}|\Omega) &= \frac{\mathrm{P}(\Omega|m_{l1}//m_{r1})\mathrm{P}(m_{l1}//m_{r1})}{\mathrm{P}(\Omega)} \qquad \text{and} \\ \mathrm{P}(m_{l1}//m_{r2}|\Omega) &= \frac{\mathrm{P}(\Omega|m_{l1}//m_{r2})\mathrm{P}(m_{l1}//m_{r2})}{\mathrm{P}(\Omega)} \end{aligned}$$

respectively where the // notation implies coupling phase. Without prior knowledge obtained by pedigree, the probability of each parental haplotype is equal to 0.5. The probability of the progeny $\mathrm{P}(\Omega)$ is equal for the two phases. Thus, the only difference in the haplotype probabilities is the probability of the progeny given the haplotype, $\mathrm{P}(\Omega|m_{l1}//m_{r1})$ or $\mathrm{P}(\Omega|m_{l1}//m_{r2})$.

This result simplifies the construction of haplotypes significantly as each haplotype can be considered separately and the haplotype with the highest probability is chosen (Haley et al., 1994 and Knott et al., 1996).

2.3.3 *Single Marker Models*

The simplest statistical model for QTL detection is the ANOVA model, assessing each marker individually (Geldermann, 1975). The model fitted contains an additive effect and (if applicable) a dominance effect. The QTL are said to be associated with the markers that are most significant and the QTL effect(s) are estimated to be those based on the marker inheritance classes. These estimates will be biased depending on how far the QTL lies from the marker. The additive effect for the marker under consideration is $a(1-2r)$ where a is the QTL additive effect and r is the recombination fraction between QTL and marker. This result is derived using Table 2.1 as half the difference between the marker homozygotes. The bias in the genetic parameters derived from the marker inheritance class means increases with the recombination fraction. The residual variance will not be homogeneous if there is dominance present in the population (Asins & Carbonell, 1988), see Table 2.1. The marker inheritance class means are estimated with differing precision if dominance is present.

To overcome the biased genetic effect estimate from single marker ANOVA models, Weller (1986) proposed a method of estimation that incorporates the recombination between the marker and the QTL. Weller (1986) realised that the markers were only proxies for the QTL and hence are only estimates for them. To overcome this he proposed that the distribution of the trait values within a marker allele class is a mixture distribution. The component distributions to be mixed over are the distributions of the three QTL classes and the mixing probabilities are based on the recombinations needed to realise the QTL class given the marker class. For the trait distributions of the low line homozygotes $f_{q_1q_1}(\cdot)$, the heterozygotes $f_{q_1q_2}(\cdot)$ and the high line homozygotes $f_{q_2q_2}(\cdot)$ the mixture distributions are (Weller, 1986)

$$\begin{aligned} f(y_i) &= (1-r)^2 f_{q_1q_1}(y_i) + 2r(1-r) f_{q_1q_2}(y_i) + r^2 f_{q_2q_2}(y_i) \\ f(y_j) &= r(1-r) f_{q_1q_1}(y_j) + 2(1-2r(1-r)) f_{q_1q_2}(y_j) + r(1-r) f_{q_2q_2}(y_j) \\ f(y_k) &= r^2 f_{q_1q_1}(y_k) + 2r(1-r) f_{q_1q_2}(y_k) + (1-r)^2 f_{q_2q_2}(y_k) \end{aligned}$$

where the first and third distribution are for the marker homozygotes, the second is for the marker heterozygotes and r is the recombination fraction. Typically the functional form of the QTL classes' distributions are assumed to be normal. Weller (1986) estimated different variances for the three QTL classes' distributions, but they may be assumed equal if appropriate (e.g. Lander & Botstein, 1989). With unequal variances, the trait density has seven parameters, three location parameters, three variances and the recombination fraction. There are only six observed pieces of information, the trait means and variances of each unknown marker type class. Hence, optimisation of this density is somewhat of an art. Weller (1986) suggested solving for the genetic parameters based mostly on the homozygote classes and then using these for estimation for the other parameters.

Table 2.1: Expected recombination frequencies for the marker and QTL inheritance classes, and the marker inheritance classes' expected trait values and variances. The recombination fraction is r, the additive QTL effect is a and the QTL dominance effect is d.

Marker Type	QTL Allele Type			Marker Inheritance Class	
	q_1q_1	q_1q_2	q_2q_2	Expected Value	Variance[1]
m_1m_1	$(1-r)^2$	$2r(1-r)$	r^2	$a(1-2r)+2dr(1-r)$	$2a^2r(1-2r)+2d^2rs-4adr(1-3r+2r^2)$
m_1m_2	$r(1-r)$	$(1-2r+2r^2)$	$r(1-r)$	$d(1-2r+2r^2)$	$2a^2r(1-r)+2d^2rs$
m_2m_2	r^2	$2r(1-r)$	$(1-r)^2$	$-a(1-2r)+2dr(1-r)$	$2a^2r(1-2r)+2d^2rs+4adr(1-3r+2r^2)$
QTL Class Mean	a	d	$-a$		

[1] $s = 1-3r+4r^2-2r^3$. Table adapted from Asins & Carbonell (1988)

2.3.4 *Interval Mapping*

Neither of the single marker methods, ANOVA or the mixture approach, takes into account the association between the markers. Ignoring the quantitative trait, the ordering of the markers and the genetic distances can be determined by considering the recombination fractions, creating a linkage map. Each linkage group in the map corresponds to a chromosome, and each marker is located on a chromosome. The task of associating the quantitative trait to positions on the genome can then be re-stated as positioning the QTL to locations on the marker map, with each position placed in between known markers, in particular those flanking the putative position.

The flanking markers, considered as a pair, can then be used to quantify the association between the genomic region and the trait. This process is called interval mapping (IM) and was first proposed by Lander & Botstein (1989). There are two estimation methods commonly used for IM: the mixture approach, an extension of the mixture likelihood single marker analysis (Weller, 1986); and a regression approach, a simplification of the mixture approach. The mixture approach is more computationally demanding and also slightly more powerful (Xu, 1995 and Kao, 2000) however, the estimates of the genetic effects are almost identical (Haley & Knott, 1992; Xu, 1995; and Kao, 2000). The regression approach is also more flexible due to reduced computation.

In both methods the putative QTL are considered to be significant if the log-likelihood ratio, or a function of it, reaches a threshold value. In mixture interval mapping the log odds of the difference (LOD) statistic is often used. The threshold for a significant LOD score value can be difficult to choose as it needs to take into account multiple testing of many non-independent tests. A common solution is to use an arbitrary cut-off, often chosen to be a LOD score of 3. Theoretical solutions that make various assumptions about the data are available (for example Lander & Botstein, 1989; Rebaï et al., 1994; Rebaï et al., 1995; Lander & Kruglyak, 1995; and Piepho, 2001). Empirical thresholds, based on permutation methods, have also been proposed (Churchill & Doerge, 1994 and Doerge & Churchill, 1996) and are common for simple models but are often considered impractical for complex models due to multiple fitting of the model. The theoretical and empirical thresholds have been shown to produce similar results, at least for one of the theoretical adjustments (Doerge & Rebaï, 1996).

The interval mapping approach has been shown to produce a more powerful detection method than the single marker methods when the marker map is known (Rebaï et al., 1995). In situations where the marker map is estimated poorly, the usefulness of IM could be questioned. However, interval mapping seems to be robust to incorrectly estimated distances between the markers (Dodds et al., 2004). Improper ordering of the markers is likely to have a much more drastic effect on interval mapping.

Mixture Interval Mapping

This approach is similar to the single marker mapping of Weller (1986). However, two markers are used to calculate the mixing probabilities for the trait distributions. Also, the variances within the trait distributions for each QTL inheritance class are assumed equal. Typically, the QTL inheritance type for an individual at a particular putative location is regarded as missing information (Lander & Botstein, 1989; Jansen, 1992; and Carbonell et al., 1992). The expectation-maximisation (EM) algorithm can be used to maximise the mixture likelihood.

The likelihood for individual i with unknown QTL inheritance type from an F_2 population is

$$L_i(\mu, a, \sigma^2) = p_i(q_1q_1)f_{q_1q_1}(y_i) + p_i(q_1q_2)f_{q_1q_2}(y_i) + p_i(q_2q_2)f_{q_2q_2}(y_i)$$

where $p_i(q_1q_1)$, $p_i(q_1q_2)$ and $p_i(q_2q_2)$ are the probability for the i^{th} individual of the QTL having types q_1q_1, q_1q_2 and q_2q_2 respectively, and $f_{q_1q_1}$, $f_{q_1q_2}$ and $f_{q_2q_2}$ are the distributions for the q_1q_1, q_1q_2 and q_2q_2 QTL classes.

The expectation step is to estimate for each individual the probability distribution p_i of the QTL classes given the trait values and the marker values. In the mixture setting this is calculating the expected value of the Bayesian posterior estimate for the QTL inheritance class conditional on the markers' inheritance classes and trait values. That is, for individual i and QTL alleles q_j and q_k, the mixing probability is (Jansen, 1992)

$$p_i(q_jq_k|y_i, \boldsymbol{M}_i) = \frac{p_i(q_jq_k|\boldsymbol{M}_i)f(y_i|q_jq_k)}{f(y_i|\boldsymbol{M}_i)} \tag{2.3.1}$$

where y_i is the measured trait for the i^{th} individual and $\boldsymbol{M}_i$ are the flanking marker types for the i^{th} individual. This form shows that the mixture IM approach takes the trait values into account when estimating the QTL inheritance types.

The maximisation step can be carried out using regression or generalised linear models for non-normal data (Jansen, 1992 and Jansen, 1993b). This process involves expanding the data set, once for each putative QTL inheritance class, using the mixing probabilities from (2.3.1) as weights and then modelling the trait values on this expanded data set.

A particular position within an interval need not be considered. Rather, consideration can be given to each interval in turn and the recombination between QTL and flanking markers can be estimated (Jansen, 1992). This estimation process is lengthy and has the potential to fail to reach convergence or to converge to a local maximum (Jansen & Stam, 1994). These authors suggest that profiling over the recombination fraction within an interval is a better solution. In practice, a set of putative QTL positions are identified and then assessed for association (Lander & Botstein, 1989 and Jansen & Stam, 1994). This estimation process will mean that many models will be fitted, one for each putative position.

Regression Interval Mapping

The mixture IM method can be computationally demanding and hence restrictive for the type and complexity of models that can be fitted. It appears that the (posterior) probability (2.3.1) of the QTL class given the marker data and the quantitative trait is dominated by the probability of the QTL classes conditional on the marker data ignoring the trait measurements. That is, there is more information about the QTL classes in the marker data than there is in the trait data (Xu, 1995). This observation implies that the posterior probability may be well approximated by the conditional probability ignoring the trait data (Kao, 2000). This method was originally proposed by Haley & Knott (1992) and Martinez & Curnow (1992). Unlike mixture IM, estimation needs to occur only once at each putative location and does not need iteration on an expanded data set.

The explanatory variables, functions of the expected conditional probabilities, are calculated where the specified putative QTL lie on the assumed marker map. In particular they are based on the probability of inheriting the different QTL alleles. For a pair of flanking markers M_l and M_r with inheritance alleles $\{m_{l1}, m_{l2}\}$ and $\{m_{r1}, m_{r2}\}$ for the left and right flanking markers respectively, the expected frequencies of the marker-QTL-marker haplotype are given in Table 2.2. The expected recombinations between the markers are also given in this table.

The expected conditional probabilities are the ratio of the expected frequency of the QTL inheritance type of interest and the expected frequency of the marker inheritance types. The explanatory variables in the linear model are formed by taking the difference of the two expectations for being homozygous and the expected conditional probability of being heterozygous. These estimate twice the additive effect ($2a$) and the dominance effect (d) respectively. The explanatory variables are given explicitly in Table 2.3 and are based on Table 2.2.

Table 2.2: Expected frequencies in an F_2 cross for putative QTL inheritance classes conditional on the marker inheritance type. The last column is the expected recombination fraction between the flanking marker pair. The recombination between markers is r, left marker and QTL r_l, and right marker and QTL r_r.

Marker Type		QTL Inheritance Class Recombination Frequency			Marker Recombination Frequency
M_l	M_r	q_1q_1	q_1q_2	q_2q_2	
$m_{l1}m_{l1}$	$m_{r1}m_{r1}$	$(1-r_l)^2(1-r_r)^2/4$	$2r_l(1-r_l)r_r(1-r_r)/4$	$r_l^2r_r^2/4$	$(1-r)^2/4$
$m_{l1}m_{l1}$	$m_{r1}m_{r2}$	$(1-r_l)^2r_r(1-r_r)/4$	$[r_l(1-r_l)(1-r_r)^2+r_l(1-r_l)r_r^2]/4$	$r_l^2r_r(1-r_r)/4$	$r(1-r)/4$
$m_{l1}m_{l1}$	$m_{r2}m_{r2}$	$(1-r_l)^2r_r^2/4$	$2r_l(1-r_l)r_r(1-r_r)/4$	$r_l^2(1-r_r)^2/4$	$r^2/4$
$m_{l1}m_{l2}$	$m_{r1}m_{r1}$	$r_l(1-r_l)(1-r_r)^2/4$	$[(1-r_l)^2r_r(1-r_r)+r_l^2r_r(1-r_r)]/4$	$r_l(1-r_r)r_r^2/4$	$r(1-r)/4$
$m_{l1}m_{l2}$	$m_{r1}m_{r2}$	$r_l(1-r_r)r_r(1-r_r)/4$	$[(1-r_l)^2(1-r_r)^2+r_l^2(1-r_r)^2+(1-r_l)^2r_r^2+r_l^2r_r^2]/4$	$r_l(1-r_r)r_r(1-r_r)/4$	$(r^2+(1-r)^2)/4$
$m_{l1}m_{l2}$	$m_{r2}m_{r2}$	$r_l(1-r_r)r_r^2/4$	$([1-r_l)^2r_r(1-r_r)+r_l^2r_r(1-r_r)]/4$	$r_l(1-r_l)(1-r_r)^2/4$	$r(1-r)/4$
$m_{l2}m_{l2}$	$m_{r1}m_{r1}$	$r_l^2(1-r_r)^2/4$	$2r_l(1-r_l)r_r(1-r_r)/4$	$(1-r_l)^2r_r^2/4$	$r^2/4$
$m_{l2}m_{l2}$	$m_{r1}m_{r2}$	$r_l^2r_r(1-r_r)/4$	$[r_l(1-r_l)(1-r_r)^2+r_l(1-r_l)r_r^2]/4$	$(1-r_l)^2r_r(1-r_r)/4$	$r(1-r)/4$
$m_{l2}m_{l2}$	$m_{r2}m_{r2}$	$r_l^2r_r^2/4$	$2r_l(1-r_l)r_r(1-r_r)/4$	$(1-r_l)^2(1-r_r)^2/4$	$(1-r)^2/4$

Table 2.3: Explanatory variables for regression mapping in an F$_2$ population. The recombination between markers is r, left marker and QTL r_l, and right marker and QTL r_r.

Marker Type		Explanatory Variables	
M_l	M_r	$2a$	d
$m_{l1}m_{l1}$	$m_{r1}m_{r1}$	$[(1-r_l)^2(1-r_r)^2-r_l^2r_r^2]/(1-r)^2$	$[2r_l(1-r_r)r_2(1-r_2)]/(1-r)$
$m_{l1}m_{l1}$	$m_{r1}m_{r2}$	$[(1-r_l)^2r_r(1-r_r)-r_l^2r_r(1-r_r)]/r(1-r)$	$[r_l(1-r_l)(1-r_r)^2+r_l(1-r_l)r_r^2]/r(1-r)$
$m_{l1}m_{l1}$	$m_{r2}m_{r2}$	$[(1-r_l)^2r_r^2-r_l^2(1-r_r)^2]/r^2$	$[2r_l(1-r_l)r_r(1-r_r)]/r^2$
$m_{l1}m_{l2}$	$m_{r1}m_{r1}$	$[r_l(1-r_l)(1-r_r)^2-r_l(1-r_r)r_r^2]/r(1-r)$	$[(1-r_l)^2r_r(1-r_r)+r_l^2r_r(1-r_r)]/r(1-r)$
$m_{l1}m_{l2}$	$m_{r1}m_{r2}$	0	$[(1-r_l)^2(1-r_r)^2+r_l^2(1-r_r)^2+(1-r_l)^2r_r^2+r_l^2r_r^2]/[r^2+(1-r)^2]$
$m_{l1}m_{l2}$	$m_{r2}m_{r2}$	$[r_l(1-r_r)r_r^2-r_l(1-r_l)(1-r_r)^2]/r(1-r)$	$[(1-r_l)^2r_r(1-r_r)+r_l^2r_r(1-r_r)]/(1-r)^2]$
$m_{l2}m_{l2}$	$m_{r1}m_{r1}$	$[r_l^2(1-r_r)^2-(1-r_l)^2r_r^2]/r^2$	$[2r_l(1-r_l)r_r(1-r_r)]/r^2$
$m_{l2}m_{l2}$	$m_{r1}m_{r2}$	$[r_l^2r_r(1-r_r)-(1-r_l)^2r_r(1-r_r)]/r(1-r)$	$[r_l(1-r_l)(1-r_r)^2+r_l(1-r_l)r_r^2]/r(1-r)$
$m_{l2}m_{l2}$	$m_{r2}m_{r2}$	$[r_l^2r_r^2-(1-r_l)^2(1-r_r)^2]/(1-r)^2$	$[2r_l(1-r_l)r_r(1-r_r)]/(1-r)^2$

The computation time for the regression interval mapping can be further reduced when models with only additive effects are considered. Many putative positions within an interval do not have to be individually assessed. Instead the position within the interval can be shown to be a function of the effects of the flanking markers (Whittaker et al., 1996). Consider the regression interval mapping model

$$\begin{aligned} \mathrm{E}\,(y) &= \beta_0 + \beta_1 \mathrm{E}\,(q|m_i m_{i+1}, r_l) \\ &= \beta_0 + a\delta m_i + a\rho m_{i+1} \\ &= \beta_0 + \beta_1^* m_i + \beta_2^* m_{i+1} \end{aligned}$$

where $\delta = \mathrm{E}\,(q|m_i = 1, m_{i+1} = -1, r_l)$ and $\rho = \mathrm{E}\,(q|m_i = -1, m_{i+1} = 1, r_l)$. The single coefficient in the simple regression model has been replaced by two coefficients (β_1 and β_2). These two coefficients can be equated to the two unknown QTL parameters, the additive effect a and the position r_l giving

$$r_l = 0.5\left(1 - \sqrt{1 - \frac{4\beta_2^* r(1-r)}{\beta_2^* + \beta_1^*(1-2r)}}\right)$$

and

$$a^2 = \frac{(\beta_1^* + (1-2r)\beta_2^*)(\beta_2^* + (1-2r)\beta_1^*)}{1-2r}$$

where r is the recombination fraction between the two markers.

Dominance terms can be incorporated into this model by estimating this effect at the estimated position (Whittaker et al., 1996). However, this may be an inefficient method to incorporate dominance as it is then only estimated at genomic positions where the additive effect is high.

One possible drawback of the regression approach using flanking markers to estimate the QTL effect and position within an interval is collinearity between the flanking markers. Flanking markers in tight linkage may cause the columns of the model's design matrix to be numerically dependent. If this occurs then the results may be unpredictable. The problem is expected to be exacerbated when more than one QTL is fitted on a chromosome in a multiple QTL model.

A criticism of regression interval mapping, irrespective of estimation procedure, is that the residual variance is over-estimated (Xu, 1995). Consider the variance of the regression model estimating only one additive effect

$$\begin{aligned} \sigma_e^2 = \mathrm{var}\,(y_j) &= \mathrm{var}\,(\beta_0 + \beta_a \mathrm{E}\,(q_j|m_{j,i} m_{j,i+1}, r_l) + e_j) \\ &= \beta_a^2 \mathrm{var}\,(\mathrm{E}\,(q_j|m_{j,i} m_{j,i+1}, r_l)) + \mathrm{var}\,(e_j) \qquad (2.3.2) \\ &\geq \mathrm{var}\,(e_j) = \sigma_\epsilon^2. \end{aligned}$$

The estimated variance σ_e^2 is a function of the true residual variance σ_ϵ^2, the additive effect β_a and the conditional probability of QTL inheritance class. This implies that the regression approach will not be as powerful as a model that incorporates the heterogeneity, such as the mixture IM approach (Xu, 1995; Xu, 1998a; and Xu, 1998b).

The regression approach is a first moment approximation. The biased residual variance can be reduced by approximating the distribution of the trait given the marker inheritance types by first and second moments. To achieve this Xu (1998a) and Xu (1998b) suggested using an iteratively re-weighted least squares (IRWLS) approach. This is an iterative procedure that uses a weighting matrix for each observation. The weights are based on the previous iteration's estimates of the residual variance, additive and dominance genetic effects and the variances of the expectation of the QTL inheritance class given the flanking markers. If $\boldsymbol{W}$ is the weight matrix at a given iteration then the weight for individual j is

$$W_{jj} = \frac{\beta_a^2}{\sigma_e^2} var(z_j|\boldsymbol{m}) + \frac{\beta_d^2}{\sigma_e^2} var(w_j|\boldsymbol{m}) + \frac{2\beta_a\beta_d}{\sigma_e^2} cov(z_j w_j|\boldsymbol{m}) + 1$$

where β_a is the additive effect, β_d is the dominance effect, $z_j = \pm 1$, $w_j = 0$ if the individual is homozygous, $z_j = 0$, $w_j = 1$ if it is heterozygous and σ_e^2 is the residual variance.

Like regression interval mapping, the IRWLS interval mapping produces unbiased estimates of the genetic effects. However, the residual variance is not inflated and hence should give a more powerful test (Xu, 1998a and Xu, 1998b). More computation is required than the regression method but assuming that good starting values can be obtained, convergence may not take more than a small number of iterations.

2.4 Multiple QTL methods

All the methods considered thus far examine the association of a particular locus on the genome ignoring all the other loci. Some of these other loci may be QTL and hence they may explain variation in the trait. Ignoring the other QTL may lead to misleading inference (e.g. Miller, 2002).

There have been a number of methods proposed that allow for multiple QTL. They can loosely be divided into three types: those extending the single marker models by fitting multiple markers; those using a combination of single interval mapping and marker models by including marker explanatory variable in a model when conducting interval mapping; and those extending interval mapping by considering multiple intervals simultaneously.

2.4.1 *Multiple Marker Models*

These models are an extension of the single marker models. However, instead of fitting a separate model for each and every marker, a single model is fitted with either all the markers present (e.g. Whittaker et al., 2000; Gianola et al., 2003; and Xu, 2003) or a selected subset of them (e.g. Broman & Speed, 2002).

In practice it is not possible, nor desirable to include all the markers scored into a single model as fixed effects. Often there are more markers than there are individual observations so the fixed effect estimates are not unique. The marker variables are highly correlated if they are on the same chromosome, potentially producing a system of estimating equations that is ill-conditioned and hence estimates that are highly variable - an undesirable situation.

To overcome these problems a forward moving selection procedure such as forward selection or stepwise selection is often employed. These methods fit a series of models, starting from a model with no marker variables. The complexity of the model is increased until all the remaining marker variables in the model are not significantly related to the outcome.

Ridge regression (Hoerl & Kennard, 1970a) was proposed by Whittaker et al. (2000) for the characterisation of the genetic merit of a set of markers. The most important of these will be those that are close, in genomic position, to the QTL. These provide proxies for the QTL.

Ridge regression can be specified as a random regression where the marker effects are identically and independently distributed as normal variates (Lindley & Smith, 1972). If the markers are assumed not to be QTL then this random effects model does not take into account all the information known about the genetical system. In particular it assumes that all the marker effects are independent. Gianola et al. (2003) reformulated the ridge regression model using the random normal specification. They allowed for the correlation by drawing the random effects from a multivariate normal distribution with an auto-regressive, or gaussian correlation structure. If the markers are equally spaced then the marker variables have an auto-regressive structure. However, this does not imply that the marker effects also have the same structure.

Another method based on ridge regression was proposed by Xu (2003). This method assumes that each marker effect had a normal distribution, however they have their own variance, further, the variance is inversely proportional to the marker effect. Xu (2003) estimated the model using Markov chain monte carlo (MCMC) methods as, analytical expressions are difficult or impossible to obtain and solve. However, as noted by ter Braak et al. (2005) the model originally proposed by Xu (2003) will produce improper posterior distributions. ter Braak et al. (2005) also suggest a different sampling method for MCMC that should explore the posterior distribution more thoroughly. The model presented in Xu (2003) has the attractive feature that the important markers are observations from a distribution with larger variance than the unimportant ones. This means that the important effects will be estimated with less shrinkage than the unimportant effects.

For any given marker, the size of its effect is dependent on the genetic distance between it and a QTL. In genomic areas where there are many markers, this diminution of effect size is unlikely to cause problems. However, if the marker map doesn't contain (approximately) equally spaced markers, then the diminution of the genetic effect will be disproportional. The markers will have different sized effects depending on whether they come from a marker rich or poor genomic area. This problem has been largely overlooked. However, in a Bayesian setting the inclusion of a marker may be assigned a higher probability if it comes from a marker poor genomic region. This is incorporated in the model averaging method of Ball (2001).

2.4.2 *Composite Interval Mapping*

While a putative position is being examined by the interval mapping method, other genetic effects can be approximately modelled by including markers that are located in proximity to other QTL. This method is called composite interval mapping (CIM) (Zeng, 1993 and Zeng, 1994) or multiple QTL mapping (MQM) (Jansen, 1993a and Jansen & Stam, 1994). Selection of the markers for inclusion into the model is a problem almost identical to the multiple marker regression problem discussed previously and can be solved by forward selection or stepwise model building.

The power to detect QTL using CIM is greater than using simple interval mapping (Zeng, 1994 and Jansen & Stam, 1994). In fact, for moderate to high marker densities the power to detect QTL using CIM is approximately the same as using the QTL not under consideration themselves (Jansen, 1994). This implies that the inclusion of markers as proxies for QTL performs well.

Initially CIM was proposed for estimation via the mixture interval mapping technique. However, some practitioners use CIM with regression interval mapping. The residual variance for the regression mapping method (Section 2.3.4) is composed of two parts, the true residual variance and the extra variance obtained from not considering the estimation problem as a mixture. Xu (1998a) warns against using regression interval mapping for CIM as the reduction in true residual variance might mean that the estimated residual variance is dominated by the component due to ignoring the mixture structure. In place of regression interval mapping the iteratively re-weighted least squares method or the mixture method could be used if this is considered a sufficient problem. However, the benefit of multiple QTL methods is likely to out weigh the inflation of variance.

Like the multiple marker models, the effect size of the non-flanking markers contained in a CIM model is dependent on distance of the QTL to the marker. For CIM this means that test statistics conditioning on markers acting as proxies for QTL in genomic regions containing few markers may not be comparable to those with markers acting as proxies in areas with many QTL (Zeng et al., 1999). The markers in the different areas will explain different amounts of residual variation.

In practice, there seems to be little difference between fitting the QTL model using CIM (with forward selection for the marker co-variables) and fitting a marker model using forward selection (Broman & Speed, 2002). This may be because, in dense or moderately dense marker maps, the marker variables themselves provide good proxies to the QTL.

The markers that are not flanking the interval of interest could be included into the model as constrained variables, potentially as ridge regression variables (Boer et al., 2002). This method removes the need to select any markers in preference for any of the others and enables all the additive genetic effects to be accounted for. Additive digenic epistatic terms can also be accounted for by including these as ridge regression parameters, potentially with a different constraint parameter from the additive main effects (Boer et al., 2002).

Testing thresholds for presence of a QTL at the putative location using CIM can be obtained theoretically by extending those for simple interval mapping. Specifically, if it is assumed that there is an infinite residual degrees of freedom, then the test statistic will follow a distribution that is between a χ^2_1 and a χ^2_2 (Jansen, 1994). Accounting for multiple testing uses the same adjustments used for simple interval mapping. Permutation thresholds can be extended to the CIM setting (Doerge & Churchill, 1996). These involve either permuting on the residuals of the model containing only the marker effects or conditioning on the marker classes of all the markers in the model.

2.4.3 *Multiple Interval Mapping*

The idea of mapping multiple QTL using multiple intervals was first introduced at the inception of interval mapping (Lander & Botstein, 1989). Commonly with mixture estimation, a forward selection procedure is used for identifying the model. However, if there is more than one QTL on the same chromosome then it will be difficult to distinguish the QTL with this procedure as both QTL will appear as a single *ghost* QTL (Haley & Knott, 1992 and Martinez & Curnow, 1992). In these cases it may be beneficial to use a two dimensional search routine to simultaneously find the two QTL on a chromosome, if there are indeed two (Haley & Knott, 1992). This two dimensional procedure is akin to finding the best model via the selection method of best subsets of size 2. This can be computationally demanding if mixture interval mapping is used for estimation and hence regression interval mapping may be preferred.

The process of fitting multiple intervals, using estimation by the mixture methods, has been named multiple interval mapping (MIM) by Kao et al. (1999). If mixture interval mapping is used then the number of mixing classes increases, by about two fold for backcross populations and about three fold for F_2 populations, each time a QTL is added to the model (Kao et al., 1999). To overcome this difficulty the estimation can be simplified by only including a small number of important mixing classes into the estimation process (Kao et al., 1999). Alternatively, a stochastic element can be introduced into the estimation procedure as was outlined in the monte carlo EM estimation method for QTL analysis by Jansen (1996). If all the mixing class probabilities in the MIM model are close to zero or one, then the estimates will be almost identical to the regression mapping approach.

A selection method that uses the inherent structure of the genetic data was presented by Piepho & Gauch (2001). The marker variables (and QTL variables if known) will be stochastically independent if they lie on different chromosomes. This implies that selection can occur sequentially for each chromosome in turn using the identified model for the other chromosomes. Piepho & Gauch (2001) use the flanking marker regression method of Whittaker et al. (1996) to estimate the QTL positions within the interval once it has been identified. This reduces the computational cost for a high dimensional search.

2.5 Bayesian Methods

Up to this point, this review has largely ignored Bayesian methods. The notable exceptions are the all marker models of Ball (2001), Gianola et al. (2003), Xu (2003) and ter Braak et al. (2005). Here, this omission is rectified by providing a brief review of some of the key ideas. This section is not comprehensive, as there is a substantial amount of literature available. Furthermore, the Bayesian paradigm is not employed in this thesis.

Many of the methods that will be described were developed for outbreeding species where there is a known pedigree structure (common in species such as dairy cattle). Consequently, many of the proposed models include a polygenic random effect for each animal. The covariance of these random effects is defined by the numerator relationship matrix. The inclusion of these effects in the model suggests that many of the following methods are an extension of the animal model which is frequently used in non-molecular quantitative genetics. Since the motivating data set for this thesis does not contain pedigree information about the parental populations, explicit mention of these terms in the following is not made. Only the portions of the models relevant to experimental crosses are presented.

2.5.1 *Single Marker Models*

The development of Bayesian QTL mapping methods has proceeded along a similar path to the non-Bayesian methods. In particular, early Bayesian methods were based on single marker models, similar to those in Section 2.3.3 (Hoeschele & VanRaden, 1993a; Hoeschele & VanRaden, 1993b; Thaller & Hoeschele, 1996a; Thaller & Hoeschele, 1996b). The underlying genetic model in these papers is similar to that presented in Weller (1986). However, Hoeschele & VanRaden (1993a) argue that sensible prior information is available for the QTL mapping problem and hence the Bayesian approach could be considered more appropriate. Another potential benefit of these models over the non-Bayesian methods is that variation in the nuisance parameters (e.g. the recombination fraction) can be accounted for by integration of the joint posterior distribution.

2.5.2 *Incorporating Multiple Markers*

The Bayesian single marker model, like the corresponding non-Bayesian model, can be extended using multiple markers. The most direct extension is to consider the QTL to be associated with a linkage group rather than just a single marker (Uimari et al., 1996; Satagopan et al., 1996). Such analyses will provide information about the linkage status of the QTL with the linkage group. Also, the marker map and QTL position can be estimated, up to a predefined order, using the marker *and* the trait values (Uimari et al., 1996). Similar to composite interval mapping, markers acting as proxies (cofactors) to QTL on different chromosomes can be included into the model to remove their genetic effect and hence reduce residual variance (e.g. Satagopan et al., 1996).

2.5.3 *Outbred Line Crosses*

As mentioned earlier in this chapter, QTL mapping in outbred line crosses requires special consideration. The phases of the parents' linked markers are unknown. This was addressed for single QTL models in half-sib designs by George et al. (2000). These authors estimate the additive and dominance genetic effects over a number of unrelated families. Estimation of the QTL location within the marker map (whose genetic distances were estimated during the analysis) was performed by using the reversible jump Markov chain Monte Carlo algorithm (RJMCMC; Green, 1995). The estimation problem can be posed as a variable selection problem where the variables to be selected are the intervals' putative QTL.

2.5.4 *Multiple QTL Models*

The single QTL models based on multiple markers are easily extended to models that include multiple QTL (e.g. Satagopan et al., 1996; Uimari & Hoeschele, 1997). The method presented in Uimari & Hoeschele (1997) is a 2 QTL version of their previous work (Uimari et al., 1996). For any given chromosome this method assesses the hypotheses that there are: 1) no QTL linked to the chromosome; 2) 1 QTL linked and one unlinked; and 3) two QTL linked to the chromosome. While these hypotheses generalise those in their previous paper, they are still quite limiting as there have to be exactly two QTL affecting the quantitative trait. The method of Satagopan et al. (1996) avoids such problems by fitting a number of models with varying assumptions about the number of QTL. The final model is chosen to be the one that substantially increases the Bayes factor (Kass & Raftery, 1995) from models with fewer QTL. Interval estimates for the position of the QTL can be obtained by using the high posterior density (HPD; e.g. Gelman et al., 1995, page 34). The distribution of these parameters may be multi-modal and hence the HPD may have disjoint intervals. This may indicate that the model does not contain a sufficient number of QTL.

The multiple QTL method of Satagopan et al. (1996) assumes a fixed number of QTL prior to undertaking analysis. In the Bayesian paradigm, this can be avoided by assuming that the number of QTL is also random. Commonly, it is assumed to be a Poisson random variable. This means that the mechanics of Bayesian variable selection, such as RJMCMC can be employed. This was first performed by Sillanpää & Arjas (1998) for inbred line crosses and subsequently for outbred populations in Sillanpää & Arjas (1999). In both these papers the genome, defined by the marker map, was split up into 'bins', which were used as units to assess the relevant posterior densities. The bins group similar samples from the Markov chain together and require fewer samples as the number of considered locations is finite.

2.5.5 *Epistatic QTL Models*

In a manner similar to the MIM method in Section 2.4.3, the Bayesian multiple QTL methods can be extended to allow pairwise (and higher order) interactions amongst QTL. Often such pairwise interactions are called digenic epistasis. This topic has been tackled in the Bayesian paradigm by Sen & Churchill (2001), Yi & Xu (2002) and Yi et al. (2003).

The method of Sen & Churchill (2001) is appealing as it does not require complex sampling methods to explore the posterior distribution. This is avoided because conditional on the QTL genotypes, the QTL effects and the trait values are independent of the marker values and the QTL position. This means that genotypes can be sampled and importance sampling (e.g. Gelman et al., 1995) can be used to evaluate the required posterior distribution. However, there are limitations of this model. In particular, most of the marginal posterior distributions are dependent on the choice of model, implying that model selection is important. Sen & Churchill (2001) proposed solution to this problem is to perform a *scan* of potential simple models, to use these to propose more complex models and then to choose a final model using Bayes factors (Kass & Raftery, 1995). This model selection method may prove to be a shortcoming of this method as it is yet unproven except for a limited number of examples.

The methods of Yi & Xu (2002) and Yi et al. (2003) use a more accepted method of model selection - the RJMCMC algorithm. Yi & Xu (2002) use RJMCMC to choose QTL for inclusion in the model irrespective of the size of its main effect, its interaction effects or both. This means that a QTL with large interaction(s) but small main effect may be inadvertently omitted from the model. This was addressed by Yi et al. (2003) by sampling the main effects and interactions separately. However, the latter method allows the inclusion of interactions without the corresponding main effects. This violates the marginality principle (Nelder, 1994).

2.5.6 *Performance of Bayesian QTL Methods*

All the analysis methods described in this section are computationally intensive. Some authors reported that their analysis took many days to compute. This is not satisfactory from a practical viewpoint. Worse, many of the estimation methods (those based on MCMC) require input from the analyst. This means that inference from an analysis is only as good as the analyst.

The only way to objectively compare analysis methods is to perform simulation studies (Broman & Speed, 2002). For some of these methods this is an almost impossible task due to the computational burden which such a study would require. This is a severe draw-back.

Chapter 3

Review: Mixed Models

3.1 Introduction

This chapter serves as a review and an introduction for the theory behind mixed models and estimation with restricted maximum likelihood (REML; Patterson & Thompson, 1971). The mixed model framework forms the basis for the methods developed in this thesis. It is not intended to be an all-encompassing exposition on the subject; for that the reader is referred to Cullis et al. (2006). Emphasis in this chapter is placed on the standard mixed model where random effects and residuals are normally distributed. The average information algorithm (Gilmour et al., 1995) is reviewed as the means for estimating the parameters. Lastly, Laplace's method for obtaining approximate marginal likelihoods and a derivative of it called partial Laplace's method (Taylor & Verbyla, 2006) are presented. These approximate likelihoods can be used for extensions of the standard mixed model when the random effects or residuals are non-normal.

This chapter makes extensive use of the statistical results in Appendix A and the vector and matrix algebra results in Appendix B.

3.2 Mixed Models

A (linear) mixed model is a model where the location effects are both fixed and random. That is

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\tau} + \boldsymbol{Z}\boldsymbol{u} + \boldsymbol{e} \tag{3.2.1}$$

where $\boldsymbol{X}$ is a $n \times p$ design matrix for the p fixed effects $\boldsymbol{\tau}$, $\boldsymbol{Z}$ is a $n \times r$ design matrix for the r random effects $\boldsymbol{u}$ and $\boldsymbol{e}$ is the $n \times 1$ vector of residuals.

The random effects $\boldsymbol{u}$ and the residuals $\boldsymbol{e}$ can come from any probability distribution, but they are commonly assumed to be independent. Another common assumption about these distributions is that they are both normal with mean zero, that is

$$\begin{bmatrix} \boldsymbol{u} \\ \boldsymbol{e} \end{bmatrix} \sim \mathrm{N}\left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \sigma^2 \begin{bmatrix} \boldsymbol{G}(\gamma) & \mathbf{0} \\ \mathbf{0} & \boldsymbol{R}(\boldsymbol{\theta}) \end{bmatrix} \right)$$

where σ^2 is the residual variance, $\boldsymbol{G}(\boldsymbol{\gamma})$ is a known $r \times r$ matrix function of the variance parameters $\boldsymbol{\gamma}$ associated with the random effects and $\boldsymbol{R}(\boldsymbol{\theta})$ is a known $n \times n$ matrix function of the variance parameters $\boldsymbol{\theta}$. The residual structures $\boldsymbol{G}(\boldsymbol{\gamma})$ and $\boldsymbol{R}(\boldsymbol{\theta})$ are correlation structures that are completely parameterised by $\boldsymbol{\gamma}$ and $\boldsymbol{\theta}$ respectively.

The expected value and the variance of the outcomes under the linear mixed model (3.2.1) are

$$\begin{aligned} \mathrm{E}\left(\boldsymbol{y}\right) &= \mathrm{E}\left(\boldsymbol{X\tau} + \boldsymbol{Zu} + \boldsymbol{e}\right) \\ &= \boldsymbol{X\tau} \qquad\qquad\qquad \text{and} \\ \mathrm{var}\left(\boldsymbol{y}\right) &= \mathrm{var}\left(\boldsymbol{X\tau} + \boldsymbol{Zu} + \boldsymbol{e}\right) \\ &= \mathrm{var}\left(\boldsymbol{Zu}\right) + \mathrm{var}\left(\boldsymbol{e}\right) \\ &= \sigma^2(\boldsymbol{ZGZ}^T + \boldsymbol{R}) \\ &= \sigma^2\boldsymbol{H}(\boldsymbol{\gamma},\boldsymbol{\theta}) \qquad\qquad \text{say.} \end{aligned}$$

If the distribution of $\boldsymbol{u}$ and $\boldsymbol{e}$ are both normal then the distribution of $\boldsymbol{y}$ will also be normal.

Neither the mean nor variance of the observed random variable $\boldsymbol{y}$ contains the unobserved random effects $\boldsymbol{u}$. Rather, the distribution of the observed data depends on the random effects only through their variances and design matrices. When estimation of parameters is carried out, it is the fixed effects and variance parameters that are estimated. If the random effects are considered interesting then they are predicted as a post-hoc procedure, conditional on knowing the variance parameters.

The estimation of the fixed effects and the variance parameters should be based on the distribution of the observed outcomes marginal to the unobserved random effects. To obtain such a distribution the joint distribution of the observed outcomes and the unobserved random effects should be integrated over with respect to the random effects. That is

$$\begin{aligned} f(\boldsymbol{y}) &= \int f(\boldsymbol{y},\boldsymbol{u})\partial\boldsymbol{u} \\ &= \int f(\boldsymbol{y}|\boldsymbol{u})f(\boldsymbol{u})\partial\boldsymbol{u} \end{aligned}$$

where the area of integration is over the support of the random vector $\boldsymbol{u}$.

Depending on the functional forms of $f(\boldsymbol{y}|\boldsymbol{u})$ and $f(\boldsymbol{u})$, this integral can be difficult to evaluate analytically. In such cases analytical approximations such as Laplace's method (Erdélyi, 1956 and de Bruijn, 1961) can be useful. Alternatively, numerical methods, such as quadrature or Markov Chain Monte Carlo, can also be employed. Fortunately, in the standard mixed model where both residual and random effects are normally distributed, the marginal distribution is also normal with functional form

$$f(\boldsymbol{y}) = \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}}\left|\boldsymbol{H}\right|^{-\frac{1}{2}}\exp\left(-\frac{1}{2\sigma^2}(\boldsymbol{y}-\boldsymbol{X\tau})^T\boldsymbol{H}^{-1}(\boldsymbol{y}-\boldsymbol{X\tau})\right).$$

This marginal likelihood for the standard normal mixed model can be maximised with respect to the fixed effects $\boldsymbol{\tau}$ and the variance parameters ratios $\boldsymbol{\gamma}$ and $\boldsymbol{\theta}$. These estimates are the maximum likelihood estimates. Maximisation for $\boldsymbol{\tau}$ yields the generalised least squares estimate

$$\hat{\boldsymbol{\tau}} = (\boldsymbol{X}^T\boldsymbol{H}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{H}^{-1}\boldsymbol{y} \tag{3.2.2}$$

which is unbiased as $\mathrm{E}\,(\hat{\boldsymbol{\tau}}) = \boldsymbol{\tau}$. From the Gauss-Markov theorem, $\hat{\boldsymbol{\tau}}$ has minimum variance of all unbiased estimates and hence is the best linear unbiased estimate (BLUE).

In a manner analogous to estimating the residual variance in linear models by maximum likelihood, the estimates of the variance parameters are biased downwards. The bias arises because estimation of the fixed effects is not taken into consideration when estimating the variance parameters. This has consequences, not only for the variance parameters themselves, but also for the estimation of the fixed effects as these are functions of the variance matrix $\boldsymbol{H}$. An unbiased method for estimating the variance parameters is obtained if the observed data are partitioned by a linear transformation. This idea was introduced by Patterson & Thompson (1971) and is called restricted maximum likelihood (REML).

3.3 Restricted Maximum Likelihood

A derivation of REML will now be given for the standard mixed model. It closely follows that given in Verbyla (1990). The observed data is transformed into two orthogonal spaces, one of dimension p that contains the fixed effects and one of dimension $(n-p)$ that contains the residuals. The partitioned data containing the residuals has a distribution free of the fixed effects but contains the variance parameters. These are estimated from this partition of the data. The fixed effects are then estimated from the partition containing both the fixed effects and the variance parameters. Details are now given.

Consider a partitioned linear transformation $\boldsymbol{W} = [\boldsymbol{W}_1\boldsymbol{W}_2]$, with $\boldsymbol{W}_1$ and $\boldsymbol{W}_2$ having dimensions $n \times p$ and $n \times (n-p)$ respectively. Further, choose $\boldsymbol{W}$ such that

$$\boldsymbol{W}_1^T\boldsymbol{X} = \boldsymbol{I}_p \qquad \text{and} \qquad \boldsymbol{W}_2^T\boldsymbol{X} = \boldsymbol{0}.$$

The first condition forces $\boldsymbol{W}_1^T\boldsymbol{y}$ to lie on the solution space for $\boldsymbol{\tau}$, that is the space spanned by $\boldsymbol{X}$. The second condition forces $\boldsymbol{W}_2^T\boldsymbol{y}$ to lie on the space orthogonal to $\boldsymbol{X}$. If $\boldsymbol{W}_1^T\boldsymbol{y}$ is chosen to be the projection of $\boldsymbol{y}$ onto the space spanned by $\boldsymbol{X}$ then $\boldsymbol{W}_2^T\boldsymbol{y}$ lies in the residual space (Seber, 1977).

The distribution of the transformed outcomes can be expressed by

$$\boldsymbol{W}\boldsymbol{y} = \begin{bmatrix} \boldsymbol{W}_1^T\boldsymbol{y} \\ \boldsymbol{W}_2^T\boldsymbol{y} \end{bmatrix} = \begin{bmatrix} \boldsymbol{y}_1 \\ \boldsymbol{y}_2 \end{bmatrix} \sim \mathrm{N}\left(\begin{bmatrix} \boldsymbol{\tau} \\ \boldsymbol{0} \end{bmatrix}, \sigma^2 \begin{bmatrix} \boldsymbol{W}_1^T\boldsymbol{H}\boldsymbol{W}_1 & \boldsymbol{W}_1^T\boldsymbol{H}\boldsymbol{W}_2 \\ \boldsymbol{W}_2^T\boldsymbol{H}\boldsymbol{W}_1 & \boldsymbol{W}_2^T\boldsymbol{H}\boldsymbol{W}_2 \end{bmatrix} \right).$$

This joint distribution shows that as far as the estimation of $\boldsymbol{\tau}$ is concerned, $\boldsymbol{y}_2$ is a covariate for $\boldsymbol{y}_1$. Hence, the appropriate distribution for the estimation of $\boldsymbol{\tau}$ is $\boldsymbol{y}_1$ conditional on $\boldsymbol{y}_2$. The conditional and marginal distributions are both normal (using Result A.1 and

Corollary A.1)

$$\begin{aligned} f(\boldsymbol{y}_1|\boldsymbol{y}_2) &\sim \mathrm{N}\left(\boldsymbol{\tau} + \boldsymbol{W}_1^T\boldsymbol{H}\boldsymbol{W}_2^T(\boldsymbol{W}_2^T\boldsymbol{H}^{-1}\boldsymbol{W}_2)^{-1}\boldsymbol{y}_2, \sigma^2(\boldsymbol{X}^T\boldsymbol{H}^{-1}\boldsymbol{X})^{-1}\right) \\ f(\boldsymbol{y}_2) &\sim \mathrm{N}\left(\boldsymbol{0}, \sigma^2\boldsymbol{W}_2^T\boldsymbol{H}\boldsymbol{W}_2\right). \end{aligned} \tag{3.3.1}$$

The distribution of $\boldsymbol{y}_2$ does not contain the fixed effects and hence it should be used for estimation of the variance parameters.

The marginal distribution in (3.3.1) contains the arbitrary transformation $\boldsymbol{W}$. However, for the standard mixed model, the elements of the functional form of the normal distribution can be expressed in terms free of this transformation (see Theorem A.2). The marginal log likelihood based on $\boldsymbol{y}_2$ is the restricted likelihood used in REML estimation. Ignoring constants it is

$$\ell_r = -\frac{1}{2}\left((n-p)\log\sigma^2 + \log|\boldsymbol{H}| + \log \boldsymbol{X}^T\boldsymbol{H}\boldsymbol{X} + \frac{\boldsymbol{y}^T\boldsymbol{P}\boldsymbol{y}}{\sigma^2}\right) \tag{3.3.2}$$

where $\boldsymbol{P} = \boldsymbol{H}^{-1} - \boldsymbol{H}^{-1}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{H}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{H}^{-1}$. The log-likelihood for estimating the variance parameters does not depend on the arbitrary transformation used to derive it.

3.4 Estimation of Fixed Effects and Variance Parameters

3.4.1 *Fixed Effects*

The conditional log-likelihood of $\boldsymbol{y}_1$ given $\boldsymbol{y}_2$ is used for estimation of the fixed effects $\boldsymbol{\tau}$. From (3.3.1) the functional form of this log-likelihood, ignoring terms not involving $\boldsymbol{\tau}$, is

$$\ell_f = -\frac{1}{2\sigma^2}\left(\boldsymbol{y}_1 - (\boldsymbol{\tau} + \boldsymbol{y}^*)\right)^T\boldsymbol{X}^T\boldsymbol{H}^{-1}\boldsymbol{X}\left(\boldsymbol{y}_1 - (\boldsymbol{\tau} + \boldsymbol{y}^*)\right)$$

where $\boldsymbol{y}^* = \boldsymbol{W}_1^T\boldsymbol{H}^{-1}\boldsymbol{W}_2(\boldsymbol{W}_2^T\boldsymbol{H}^{-1}\boldsymbol{W}_2)^{-1}\boldsymbol{y}_2$.

Maximising and solving this log-likelihood with respect to $\boldsymbol{\tau}$ gives the REML estimator for the fixed effects

$$\begin{aligned} \hat{\boldsymbol{\tau}} &= \boldsymbol{y}_1 - \boldsymbol{W}_1^T\boldsymbol{H}^{-1}\boldsymbol{W}_2(\boldsymbol{W}_2^T\boldsymbol{H}^{-1}\boldsymbol{W}_2)^{-1}\boldsymbol{y}_2 \\ &= \boldsymbol{W}_1^T\left(\boldsymbol{H} - \boldsymbol{H}^{-1}\boldsymbol{W}_2(\boldsymbol{W}_2^T\boldsymbol{H}^{-1}\boldsymbol{W}_2)^{-1}\boldsymbol{W}_2^T\boldsymbol{H}\right)\boldsymbol{H}^{-1}\boldsymbol{y} \\ &= \boldsymbol{W}_1\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{H}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{H}^{-1}\boldsymbol{y} \qquad \text{from Lemma B.1} \\ &= (\boldsymbol{X}^T\boldsymbol{H}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{H}^{-1}\boldsymbol{y} \end{aligned}$$

which is the same form as the fixed effects from maximum likelihood estimation (3.2.2). However, the REML estimates will be different from the maximum likelihood estimates as the variance parameters and hence the variance matrix $\boldsymbol{H}$ will have different estimates. Like the maximum likelihood estimates, these are BLUEs for given variance parameters.

3.4.2 *Variance Parameters*

The restricted log-likelihood (3.3.2) is used for estimation of the variance parameters. It is illustrative for estimation to combine the variance parameters associated with the random effects and the residuals together, let $\boldsymbol{\eta} = (\boldsymbol{\gamma}^T, \boldsymbol{\theta}^T)^T$. The REML estimates of the variance parameters are those which solve the score equations (see Theorem A.3 for details of derivation of the score for η_i)

$$\begin{aligned} U(\sigma^2) &= \frac{\partial \ell_r}{\partial \sigma^2} \\ &= -\frac{1}{2}\left(\frac{(n-p)}{\sigma^2} - \frac{\boldsymbol{y}^T\boldsymbol{P}\boldsymbol{y}}{\sigma^4}\right) && (3.4.1) \\ U(\eta_i) &= \frac{\partial \ell_r}{\partial \eta_i} \\ &= -\frac{1}{2}\left(\text{tr}\left(\boldsymbol{P}\dot{\boldsymbol{H}}_i\right) - \frac{\boldsymbol{y}^T\boldsymbol{P}\dot{\boldsymbol{H}}_i\boldsymbol{P}\boldsymbol{y}}{\sigma^2}\right) && (3.4.2) \end{aligned}$$

where η_i is the i^{th} parameter of $\boldsymbol{\eta}$, $i = 1 \ldots n_k$, n_k is the number of variance parameters and $\dot{\boldsymbol{H}}_i$ is the derivative of $\boldsymbol{H}$ with respect to η_i.

Solving the score $U(\sigma^2)$ gives the closed form for the estimate of residual variance

$$\hat{\sigma}^2 = \frac{\boldsymbol{y}^T\boldsymbol{P}\boldsymbol{y}}{(n-p)}$$

which is the residual sums of squares divided by the degrees of freedom. In general there is no closed form for the estimates of the other variance parameters and numerical methods are usually employed to find the estimates. Commonly iterative root finding procedures, such as the Newton-Raphson (NR) or Fisher scoring (FS) methods are used. Other methods such as the EM-algorithm or variants of it are also employed. The root finding procedures update the target vector ($\boldsymbol{\eta}$) by an amount that depends on the value of the score for the previous iteration's value and a descent direction $\boldsymbol{D}$. The updates for the $(m+1)^{th}$ iteration are of the form

$$\boldsymbol{\eta}^{(m+1)} = \boldsymbol{\eta}^{(m)} - \boldsymbol{D}\left(\boldsymbol{\eta}^{(m)}\right) U\left(\boldsymbol{\eta}^{(m)}\right).$$

The descent directions $\boldsymbol{D}$ are chosen so that the updated solution is closer to the final solution than the previous solution. Common choices for $\boldsymbol{D}$ are the inverse of the Hessian of the restricted log-likelihood (the negative observed information) and its expected value (the negative expected information). These choices give the NR and FS methods respectively. The elements of the observed information matrix are derived in Theorem A.5 and are

$$\begin{aligned} \mathcal{I}_o(\sigma^2, \sigma^2) &= \frac{(n-p)}{2\sigma^4} - \frac{\boldsymbol{y}^T\boldsymbol{P}\boldsymbol{y}}{\sigma^6} \\ \mathcal{I}_o(\sigma^2, \eta_i) &= \frac{\boldsymbol{y}^T\boldsymbol{P}\dot{\boldsymbol{H}}_i\boldsymbol{P}\boldsymbol{y}}{2\sigma^4} \\ \mathcal{I}_o(\eta_i, \eta_j) &= \frac{1}{2}\text{tr}\left(\boldsymbol{P}\ddot{\boldsymbol{H}}_{ij}\right) - \frac{1}{2}\text{tr}\left(\boldsymbol{P}\dot{\boldsymbol{H}}_i\boldsymbol{P}\dot{\boldsymbol{H}}_j\right) + \frac{\boldsymbol{y}^T\boldsymbol{P}\dot{\boldsymbol{H}}_i\boldsymbol{P}\dot{\boldsymbol{H}}_j\boldsymbol{P}\boldsymbol{y}}{\sigma^2} - \frac{\boldsymbol{y}^T\boldsymbol{P}\ddot{\boldsymbol{H}}_{ij}\boldsymbol{P}\boldsymbol{y}}{2\sigma^2}. \end{aligned}$$

The elements of the expected information matrix are derived in Theorem A.6

$$\begin{aligned}\mathcal{I}_e(\sigma^2, \sigma^2) &= \frac{(n-p)}{2\sigma^4}\\ \mathcal{I}_e(\sigma^2, \eta_i) &= \frac{1}{2\sigma^2}\mathrm{tr}\left(\boldsymbol{P}\dot{\boldsymbol{H}}_i\right)\\ \mathcal{I}_e(\eta_i, \eta_j) &= \frac{1}{2}\mathrm{tr}\left(\boldsymbol{P}\dot{\boldsymbol{H}}_i\boldsymbol{P}\dot{\boldsymbol{H}}_j\right).\end{aligned}$$

Patterson & Thompson (1971) originally employed the FS method to optimise the restricted log-likelihood for the variance parameters. This can prove to be computationally demanding for moderate to large data sets as the square matrices in the trace terms have rank equal to the number of observations. The need to calculate these matrices is removed if the descent directions used are a combination (an average) of the observed and expected information (Gilmour et al., 1995). This idea is explored later when computational methods are discussed in Section 3.6.

3.5 Prediction of Random Effects

As noted earlier, the distribution of the observed outcomes $\boldsymbol{y}$ is not dependent on the unobserved random effects $\boldsymbol{u}$. However, the $\boldsymbol{u}$ do appear in the linear model (3.2.1) and may be of interest for interpretation. These effects are drawn from a known distribution, with mean zero and variance defined by the (estimated) variance parameters. For this reason, prediction of the random effects once the variance parameters are known is conceptually different from estimation of the variance components and fixed effects.

The predictions should be unbiased in distribution. That is, they have zero expected value and their variance is equal to the relevant variance component. Another desirable attribute that could reasonably be imposed is that the predictions have minimum variance of all such unbiased predictors. These requirements, along with linearity, lead to a class of predictors known as best linear unbiased predictors (BLUPs). For a discussion on the use of BLUPs see Robinson (1991). Sometimes, for example in Searle et al. (1992), linear predictors of the random effects only are called best linear predictors (BLP). No distinction is made here.

The minimum variance requirement of BLUPs implies that the distance between the unbiased predictors and the true effects is minimised amongst all choices of predictors. This leads to the BLUPs being the unbiased predictors that minimise the mean squared error (MSE) between the true and predicted effects. So

$$\begin{aligned}\mathrm{MSE} &= \mathrm{E}\left((u - \tilde{u}(\boldsymbol{y}))^2\right)\\ &= \mathrm{E}\left(\mathrm{E}\left((u - \tilde{u}(\boldsymbol{y}))^2|\boldsymbol{y}\right)\right)\\ &= \mathrm{E}\left(\mathrm{E}\left(u^2|\boldsymbol{y}\right) - 2\mathrm{E}\left(u|\boldsymbol{y}\right)\tilde{u}(\boldsymbol{y}) + \tilde{u}^2(\boldsymbol{y})\right)\\ &= \mathrm{E}\left(\mathrm{var}\left(u|\boldsymbol{y}\right) + \mathrm{E}^2\left(u|\boldsymbol{y}\right) - 2\mathrm{E}\left(u|\boldsymbol{y}\right)\tilde{u}(\boldsymbol{y}) + \tilde{\boldsymbol{u}}^2(\boldsymbol{y})\right)\\ &= \mathrm{E}\left(\mathrm{var}\left(u|\boldsymbol{y}\right)\right) + \mathrm{E}\left(\left(\mathrm{E}\left(u|\boldsymbol{y}\right) - \tilde{u}(\boldsymbol{y})\right)^2\right) \qquad (3.5.1)\end{aligned}$$

showing that MSE is minimised if and only if $\tilde{u} = \mathrm{E}\left(u|\boldsymbol{y}\right)$.

3.5.1 *Predictions to Minimise Prediction Error*

The forms for the estimates and predictions are required if they are to be easily calculated. Here, a derivation giving both the best linear unbiased estimates (BLUEs) and the best linear unbiased predictors (BLUPs) is presented. It requires knowledge of only the first and second moments of the distribution of the random effects and residuals.

Theorem 3.1. *(Cullis et al., 2006) Let $\boldsymbol{c}_1$ and $\boldsymbol{c}_2$ be $p \times 1$ and $q \times 1$ vectors respectively. Also let σ^2 and $\boldsymbol{H}$ be known. The predictor of the linear combination of fixed and random effects $\boldsymbol{c}_1^T\boldsymbol{\tau} + \boldsymbol{c}_2^T\boldsymbol{u}$ which has the minimum MSE among the class of unbiased predictors is $\boldsymbol{c}_1^T\hat{\boldsymbol{\tau}} + \boldsymbol{c}_2^T\tilde{\boldsymbol{u}}$ where*

$$\begin{aligned} \hat{\boldsymbol{\tau}} &= (\boldsymbol{X}^T\boldsymbol{H}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{H}^{-1}\boldsymbol{y} \qquad\qquad \textit{and} \\ \tilde{\boldsymbol{u}} &= \boldsymbol{G}\boldsymbol{Z}^T\boldsymbol{P}\boldsymbol{y}. \end{aligned}$$

Proof. Let $\boldsymbol{a}^T\boldsymbol{y}$ be an unbiased linear predictor of $\boldsymbol{c}_1^T\boldsymbol{\tau} + \boldsymbol{c}_2^T\boldsymbol{u}$, chosen so that its expected value matches that of the data

$$\begin{aligned} \mathrm{E}\left(\boldsymbol{a}^T\boldsymbol{y}\right) &= \mathrm{E}\left(\boldsymbol{c}_1^T\boldsymbol{\tau} + \boldsymbol{c}_2^T\boldsymbol{u}\right) \\ \boldsymbol{a}^T\boldsymbol{X}\boldsymbol{\tau} &= \mathrm{E}\left(\boldsymbol{c}_1^T\boldsymbol{\tau}\right) + \mathrm{E}\left(\boldsymbol{c}_2^T\boldsymbol{u}\right) \\ \boldsymbol{a}^T\boldsymbol{X}\boldsymbol{\tau} &= \boldsymbol{c}_1^T\boldsymbol{\tau} + \boldsymbol{0} \end{aligned}$$

giving

$$\boldsymbol{a}^T\boldsymbol{X} = \boldsymbol{c}_1^T. \tag{3.5.2}$$

This condition imposes a constraint on the optimisation problem, arising because not all values of $\boldsymbol{c}_1$ induce the correct expected value.

The MSE for the predictor $\boldsymbol{a}^T\boldsymbol{y}$ is given by

$$\begin{aligned} \mathrm{MSE} &= \mathrm{E}\left((\boldsymbol{a}^T\boldsymbol{y} - (\boldsymbol{c}_1^T\boldsymbol{\tau} + \boldsymbol{c}_2^T\boldsymbol{u}))^2\right) \\ &= \mathrm{E}\left((\boldsymbol{a}^T\boldsymbol{y} - \boldsymbol{c}_2^T\boldsymbol{u})^2\right) - \mathrm{E}\left(2(\boldsymbol{a}^T\boldsymbol{y} - \boldsymbol{c}_2^T\boldsymbol{u})\boldsymbol{c}_1^T\boldsymbol{\tau}\right) + \mathrm{E}\left((\boldsymbol{c}_1^T\boldsymbol{\tau})^2\right) \\ &= \mathrm{var}\left(\boldsymbol{a}^T\boldsymbol{y} - \boldsymbol{c}_2^T\boldsymbol{u}\right) + \mathrm{E}^2\left(\boldsymbol{a}^T\boldsymbol{y} - \boldsymbol{c}_2^T\boldsymbol{u}\right) \\ &\qquad\qquad - 2\mathrm{E}\left(\boldsymbol{a}^T\boldsymbol{y} - \boldsymbol{c}_2^T\boldsymbol{u}\right)\mathrm{E}\left(\boldsymbol{c}_1^T\boldsymbol{\tau}\right) + \mathrm{E}\left((\boldsymbol{c}_1^T\boldsymbol{\tau})^2\right) \\ &= \mathrm{var}\left(\boldsymbol{a}^T\boldsymbol{y} - \boldsymbol{c}_2^T\boldsymbol{u}\right) + (\boldsymbol{a}^T\boldsymbol{X}\boldsymbol{\tau})^2 - 2\boldsymbol{a}^T\boldsymbol{X}\boldsymbol{\tau}\boldsymbol{c}_1^T\boldsymbol{\tau} + (\boldsymbol{c}_1^T\boldsymbol{\tau})^2 \\ &= \mathrm{var}\left(\boldsymbol{a}^T\boldsymbol{y} - \boldsymbol{c}_2^T\boldsymbol{u}\right) + (\boldsymbol{c}_1^T\boldsymbol{\tau})^2 - 2(\boldsymbol{c}_1^T\boldsymbol{\tau})^2 + (\boldsymbol{c}_1^T\boldsymbol{\tau})^2 \\ &= \mathrm{var}\left(\boldsymbol{a}^T\boldsymbol{y} - \boldsymbol{c}_2^T\boldsymbol{u}\right) \\ &= \sigma^2\left(\boldsymbol{a}^T\boldsymbol{H}\boldsymbol{a} + \boldsymbol{c}_2^T\boldsymbol{G}\boldsymbol{c}_2 - 2\boldsymbol{c}_2^T\boldsymbol{G}\boldsymbol{a}\right) \end{aligned}$$

To find the optimal linear combination $\boldsymbol{a}$, MSE is minimised subject to the constraint (3.5.2). This is facilitated by defining the Lagrangian $S = \mathrm{MSE} + 2\sigma^2\boldsymbol{\lambda}^T(\boldsymbol{c}_1 - \boldsymbol{X}^T\boldsymbol{a})$ where

$2\sigma^2\boldsymbol{\lambda}$ is a $p \times 1$ vector of Lagrange multipliers. Using Result B.1 the partial derivatives of S with respect to $\boldsymbol{a}$ and $\boldsymbol{\lambda}$ are

$$\begin{aligned}\frac{\partial S}{\partial \boldsymbol{a}} &= 2\sigma^2(\boldsymbol{Ha} - \boldsymbol{ZGc}_2 - \boldsymbol{X\lambda}) \qquad \text{and}\\ \frac{\partial S}{\partial \boldsymbol{\lambda}} &= 2\sigma^2(\boldsymbol{c}_1 - \boldsymbol{X}^T\boldsymbol{a}).\end{aligned}$$

These give

$$\begin{aligned}\boldsymbol{a} &= \boldsymbol{H}^{-1}(\boldsymbol{ZGc}_2 + \boldsymbol{X\lambda}) \qquad \text{and}\\ \boldsymbol{X}^T\boldsymbol{a} &= \boldsymbol{c}_1.\end{aligned} \tag{3.5.3}$$

Solving for $\boldsymbol{\lambda}$ gives

$$\boldsymbol{\lambda} = \left(\boldsymbol{X}^T\boldsymbol{H}^{-1}\boldsymbol{X}\right)^{-1}\left(\boldsymbol{c}_1 - \boldsymbol{X}^T\boldsymbol{H}^{-1}\boldsymbol{ZGc}_2\right).$$

This is substituted for $\boldsymbol{a}$ in (3.5.3) to give

$$\begin{aligned}\boldsymbol{a} &= \boldsymbol{H}^{-1}\boldsymbol{ZGc}_2 + \boldsymbol{H}^{-1}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{H}^{-1}\boldsymbol{X})^{-1}\boldsymbol{c}_1 - \boldsymbol{H}^{-1}\boldsymbol{X}(\boldsymbol{X}'^T\boldsymbol{H}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'^T\boldsymbol{H}^{-1}\boldsymbol{ZGc}_2\\ &= \left(\boldsymbol{H}^{-1} - \boldsymbol{H}^{-1}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{H}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{H}^{-1}\right)\boldsymbol{ZGc}_2 + \boldsymbol{H}^{-1}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{H}^{-1}\boldsymbol{H})^{-1}\boldsymbol{c}_1\\ &= \boldsymbol{PZGc}_2 + \boldsymbol{H}^{-1}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{H}^{-1}\boldsymbol{H})^{-1}\boldsymbol{c}_1\end{aligned}$$

evaluation of the linear combination gives

$$\begin{aligned}\boldsymbol{a}^T\boldsymbol{y} &= (\boldsymbol{PZGc}_2 + \boldsymbol{H}^{-1}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{H}^{-1}\boldsymbol{H})^{-1}\boldsymbol{c}_1)^T\boldsymbol{y}\\ &= \boldsymbol{c}_1^T(\boldsymbol{X}^T\boldsymbol{H}^{-1}\boldsymbol{H})^{-1}\boldsymbol{X}^T\boldsymbol{H}^{-1}\boldsymbol{y} + \boldsymbol{c}_2^T\boldsymbol{GZ}^T\boldsymbol{Py}\\ &= \boldsymbol{c}_1^T\hat{\boldsymbol{\tau}} + \boldsymbol{c}_2^T\tilde{\boldsymbol{u}}.\end{aligned}$$

□

3.5.2 *Mixed Model Equations*

The best predictions for the random effects are the expected values of the effects conditional on the observed outcomes (3.5.1). This suggests a different method for calculating the fixed and random effects in the standard mixed model, a method that reproduces the more general Theorem 3.1. This alternative method hinges on the fact that in the standard mixed model the mode of the so-called joint distribution of the observed outcomes and the unobserved random effects coincides with the mean of the conditional (predictive or posterior) distribution for the random variables. In the alternative method the joint distribution of the random effects and the outcomes is maximised with respect to the fixed and random effects (Henderson, 1950).

The joint distribution of the outcomes and the random effects is

$$\begin{bmatrix} \boldsymbol{y} \\ \boldsymbol{u} \end{bmatrix} \sim \mathrm{N}\left(\begin{bmatrix} \boldsymbol{X\tau} \\ \boldsymbol{0} \end{bmatrix}, \sigma^2 \begin{bmatrix} \boldsymbol{H} & \boldsymbol{ZG} \\ \boldsymbol{GZ}^T & \boldsymbol{G} \end{bmatrix} \right).$$

The conditional distribution of the outcomes given the random effects is $\mathrm{N}(\boldsymbol{X\tau}+\boldsymbol{Zu}, \sigma^2\boldsymbol{R})$ and the marginal distribution of the random effects is $\mathrm{N}(\boldsymbol{0}, \sigma^2\boldsymbol{G})$. The joint log-density for known variance parameters, ignoring terms not involving $\boldsymbol{\tau}$ nor $\boldsymbol{u}$, is

$$\begin{aligned} \log f(\boldsymbol{y}, \boldsymbol{u}) &= \log f(\boldsymbol{y}|\boldsymbol{u}) + \log f(\boldsymbol{u}) \\ &= -\frac{1}{2\sigma^2}\left((\boldsymbol{y} - \boldsymbol{X\tau} - \boldsymbol{Zu})^T \boldsymbol{R}^{-1}(\boldsymbol{y} - \boldsymbol{X\tau} - \boldsymbol{Zu}) + \boldsymbol{u}^T\boldsymbol{Gu} \right). \end{aligned}$$

Differentiation with respect to $\boldsymbol{\tau}$ and $\boldsymbol{u}$ gives the mixed model equations (MME). These are often expressed as the matrix equation

$$\begin{bmatrix} \boldsymbol{X}^T\boldsymbol{R}^{-1}\boldsymbol{X} & \boldsymbol{X}^T\boldsymbol{R}^{-1}\boldsymbol{Z} \\ \boldsymbol{Z}^T\boldsymbol{R}^{-1}\boldsymbol{X} & \boldsymbol{Z}^T\boldsymbol{R}^{-1}\boldsymbol{Z} + \boldsymbol{G}^{-1} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\tau}} \\ \tilde{\boldsymbol{u}} \end{bmatrix} = \begin{bmatrix} \boldsymbol{X}^T\boldsymbol{R}^{-1}\boldsymbol{y} \\ \boldsymbol{Z}^T\boldsymbol{R}^{-1}\boldsymbol{y} \end{bmatrix}. \tag{3.5.4}$$

Theorem 3.2. *(e.g. Cullis et al., 2006) The general BLUP solutions to the fixed effects and random effects given in Theorem 3.1 can be obtained for the standard mixed model by solving the mixed model equations.*

Proof. The approach taken is to solve the MME using Gaussian elimination. Absorb the coefficient matrix on the left hand side of (3.5.4) and the outcome vector on the right hand side of (3.5.4), giving an expression for the fixed effects free of the random effects

$$\begin{aligned} &(\boldsymbol{X}^T\boldsymbol{R}^{-1}\boldsymbol{X} - \boldsymbol{X}^T\boldsymbol{R}^{-1}\boldsymbol{Z}(\boldsymbol{Z}^T\boldsymbol{R}^{-1}\boldsymbol{Z} + \boldsymbol{G}^{-1})^{-1}\boldsymbol{Z}^T\boldsymbol{R}^{-1}\boldsymbol{X})\hat{\boldsymbol{\tau}} + \boldsymbol{0u} = \\ &\qquad \boldsymbol{X}^T\boldsymbol{R}^{-1}\boldsymbol{y} - \boldsymbol{X}^T\boldsymbol{R}^{-1}\boldsymbol{Z}(\boldsymbol{Z}^T\boldsymbol{R}^{-1}\boldsymbol{Z} + \boldsymbol{G}^{-1})^{-1}\boldsymbol{Z}^T\boldsymbol{R}^{-1}\boldsymbol{y}. \end{aligned}$$

Using Result B.2

$$\boldsymbol{X}^T\boldsymbol{H}^{-1}\boldsymbol{X}\hat{\boldsymbol{\tau}} = \boldsymbol{X}^T\boldsymbol{H}^{-1}\boldsymbol{y}$$

giving

$$\hat{\boldsymbol{\tau}} = (\boldsymbol{X}^T\boldsymbol{H}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}\boldsymbol{H}^{-1}\boldsymbol{y}.$$

Back-solve for the random effects

$$\boldsymbol{Z}^T\boldsymbol{R}^{-1}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{H}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{H}^{-1}\boldsymbol{y} + (\boldsymbol{Z}^T\boldsymbol{R}^{-1}\boldsymbol{Z} + \boldsymbol{G}^{-1})\tilde{\boldsymbol{u}} = \boldsymbol{Z}^T\boldsymbol{R}^{-1}\boldsymbol{y}$$

giving

$$\begin{aligned} \tilde{\boldsymbol{u}} &= (\boldsymbol{Z}^T\boldsymbol{R}^{-1}\boldsymbol{Z} + \boldsymbol{G}^{-1})^{-1}\boldsymbol{Z}^T\boldsymbol{R}^{-1}\boldsymbol{H}(\boldsymbol{H}^{-1} - \boldsymbol{H}^{-1}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{H}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{H}^{-1})\boldsymbol{y} \\ &= (\boldsymbol{Z}^T\boldsymbol{R}^{-1}\boldsymbol{Z} + \boldsymbol{G}^{-1})^{-1}\boldsymbol{Z}^T\boldsymbol{R}^{-1}\boldsymbol{HPy} \end{aligned}$$

and an application of Result B.2 gives

$$\tilde{\boldsymbol{u}} = \boldsymbol{GZ}^T\boldsymbol{Py}.$$

□

3.5.3 *Residuals*

A general form for the predicted residuals will be useful for computation. From Theorem 3.1 the estimated fixed and predicted random effects are known for a mixed model where the first two moments of the relevant distributions are known. The corresponding residuals are now given.

Theorem 3.3. *The residuals from the mixed model described for Theorem 3.1 where the location effects are predicted by BLUP are*

$$\tilde{e} = \boldsymbol{R}\boldsymbol{P}\boldsymbol{y}.$$

Proof. The residuals are

$$\begin{aligned}\tilde{\boldsymbol{e}} &= \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\tau}} - \boldsymbol{Z}\tilde{\boldsymbol{u}}\\ &= \boldsymbol{y} - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{H}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{H}^{-1}\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}^T\boldsymbol{P}\boldsymbol{y}\\ &= \left(\boldsymbol{H}(\boldsymbol{H}^{-1} - \boldsymbol{H}^{-1}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{H}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{H}^{-1}) - \boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}^T\boldsymbol{P}\right)\boldsymbol{y}\\ &= (\boldsymbol{H} - \boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}^T)\boldsymbol{P}\boldsymbol{y}\\ &= \boldsymbol{R}\boldsymbol{P}\boldsymbol{y}.\end{aligned}$$

□

3.6 Computation for the Standard Mixed Model

The iterative descent procedure used for solving the score equations is now defined in more detail. The standard methods of Newton-Raphson (NR) and Fisher Scoring (FS) can require heavy computation. The matrix $\boldsymbol{P}$ is required to be evaluated explicitly. It is a function of $\boldsymbol{H}^{-1}$; the inverse of a potentially large matrix.

The quadratic forms in the observed information do not require explicit evaluation of $\boldsymbol{P}$. These terms can be found numerically by using the (augmented) mixed model equations and Gaussian elimination, in particular the process of absorption. See Seber (1977) for a description for fixed effects only models and calculation of the residual sums of squares. The mixed model equations are augmented in the first column and row by working variates $\{\boldsymbol{q}_i\}$ and the first element of the matrix is $\boldsymbol{q}_i^T\boldsymbol{R}^{-1}\boldsymbol{q}_i$. Seber (1977), Gilmour et al. (1995) and Cullis et al. (2006) show that after absorption, when $\boldsymbol{q}_i = \boldsymbol{y}$, the first element is the residual sums of squares $\boldsymbol{y}^T\boldsymbol{P}\boldsymbol{y}$. The other quadratic terms can likewise be evaluated with judicious choices for the working variates.

This suggests, from a computational view, that a descent direction that contains only the quadratic forms and not the trace terms is desirable. Gilmour et al. (1995) introduced the idea of combining the observed and expected information matrices in such a manner that removes the trace terms. This combination matrix is called the average information matrix as it is approximately an average. Its elements are formed by taking an element-wise convex combination of the observed information and the expected information. The elements of the average information matrix ($\mathcal{I}_a$) are

$$\begin{aligned}
\mathcal{I}_a(\sigma^2, \sigma^2) &= \frac{1}{2}\mathcal{I}_o(\sigma^2, \sigma^2) + \frac{1}{2}\mathcal{I}_e(\sigma^2, \sigma^2) \\
&= \frac{1}{2\sigma^2}\boldsymbol{y}^T\boldsymbol{P}\boldsymbol{y} \\
\mathcal{I}_a(\sigma^2, \eta_i) &= 1\mathcal{I}_o(\sigma^2, \eta_i) + 0\mathcal{I}_e(\sigma^2, \eta_i) \\
&= \frac{1}{2\sigma^4_{\boldsymbol{H}}}\boldsymbol{y}^T\boldsymbol{P}\dot{\boldsymbol{H}}_i\boldsymbol{P}\boldsymbol{y} \\
\mathcal{I}_a(\eta_i, \eta_j) &= \frac{1}{2}\mathcal{I}_o(\eta_i, \eta_j) + \frac{1}{2}\mathcal{I}_e(\eta_i, \eta_j) \\
&= \frac{1}{2}\left[\operatorname{tr}\left(\boldsymbol{P}\ddot{\boldsymbol{H}}_{ij}\right) - \frac{1}{2}\operatorname{tr}\left(\boldsymbol{P}\dot{\boldsymbol{H}}_i\boldsymbol{P}\dot{\boldsymbol{H}}_j\right) + \frac{1}{\sigma^2}\boldsymbol{y}^T\boldsymbol{P}\dot{\boldsymbol{H}}_i\boldsymbol{P}\dot{\boldsymbol{H}}_j\boldsymbol{P}\boldsymbol{y}\right. \\
&\qquad\qquad \left. - \frac{1}{2\sigma^2}\boldsymbol{y}^T\boldsymbol{P}\ddot{\boldsymbol{H}}_{ij}\boldsymbol{P}\boldsymbol{y} + \frac{1}{2}\operatorname{tr}\left(\boldsymbol{P}\dot{\boldsymbol{H}}_i\boldsymbol{P}\dot{\boldsymbol{H}}_j\right)\right] \\
&\approx \frac{1}{2\sigma^2}\boldsymbol{y}^T\boldsymbol{P}\dot{\boldsymbol{H}}_i\boldsymbol{P}\dot{\boldsymbol{H}}_j\boldsymbol{P}\boldsymbol{y}.
\end{aligned}$$

Note that the approximation for $\mathcal{I}_a(\eta_i, \eta_j)$ is exact for models where the variance parameters enter the variance matrix linearly.

The matrix $\mathcal{I}_a$ consists only of quadratic forms. It is also a convex combination of known descent directions for the variance parameters. Implying that $\mathcal{I}_a$ is also a descent direction for the variance parameters. It is a better choice for use in descent methods as each descent requires less computation, even if it requires more descent steps to reach convergence.

3.6.1 *Working Variates*

The working variates needed to produce the quadratic forms after absorption are $\boldsymbol{q}_1$ and $\boldsymbol{q}_i$ for σ^2 and η_i respectively. These variates are

$$\begin{aligned}
\boldsymbol{q}_1 &= \boldsymbol{y} \\
\boldsymbol{q}_i &= \dot{\boldsymbol{H}}_i\boldsymbol{P}\boldsymbol{y}
\end{aligned}$$

Theorem 3.4. *The working variates $\boldsymbol{q}_1$ and $\boldsymbol{q}_i$ generate, after absorption, the required quadratic forms.*

Proof. Recall that $\boldsymbol{P}\boldsymbol{H}\boldsymbol{P} = \boldsymbol{P}$. So

$$\begin{aligned}
\boldsymbol{q}_1^T\boldsymbol{P}\boldsymbol{q}_1 &= \boldsymbol{y}^T\boldsymbol{P}\boldsymbol{y} \\
\boldsymbol{q}_1^T\boldsymbol{P}\boldsymbol{q}_i &= \boldsymbol{y}^T\boldsymbol{P}\dot{\boldsymbol{H}}_i\boldsymbol{P}\boldsymbol{y} \\
\boldsymbol{q}_i^T\boldsymbol{P}\boldsymbol{q}_i &= \boldsymbol{y}^T\boldsymbol{P}\dot{\boldsymbol{H}}_i\boldsymbol{P}\dot{\boldsymbol{H}}_j\boldsymbol{P}\boldsymbol{y}.
\end{aligned}$$

□

The working variates $\{\boldsymbol{q}_i\}$ are functions of the matrix $\boldsymbol{P}$, precisely the matrix that is demanding to compute. Fortunately, the expressions for the working variates can be expressed in a form free of this matrix.

Theorem 3.5. *The working variates* $\{\boldsymbol{q}_i\}$ *corresponding to residual variance parameters* $\{\theta_i\}$ *are given by*

$$\boldsymbol{q}_i = \dot{\boldsymbol{R}}_i \boldsymbol{R}^{-1} \tilde{\boldsymbol{e}}.$$

The working variates $\{\boldsymbol{q}_i\}$ *corresponding to a non-residual variance parameters* $\{\gamma_i\}$ *are given by*

$$\boldsymbol{q}_i = \boldsymbol{Z}_i \dot{\boldsymbol{G}}_i \boldsymbol{Z}_i^{-1} \tilde{\boldsymbol{u}}.$$

Proof. From Theorem 3.4 the working variates are $\boldsymbol{q}_i = \dot{\boldsymbol{H}}_i \boldsymbol{P}\boldsymbol{y}$. For $\boldsymbol{q}_i$ corresponding to a residual parameter $\dot{\boldsymbol{H}}_i = \dot{\boldsymbol{R}}_i$ and

$$\begin{aligned}
\boldsymbol{q}_i &= \dot{\boldsymbol{R}}_i \boldsymbol{P}\boldsymbol{y} \\
&= \dot{\boldsymbol{R}}_i \boldsymbol{R}^{-1} \boldsymbol{R}\boldsymbol{P}\boldsymbol{y} \\
&= \dot{\boldsymbol{R}}_i \boldsymbol{R}^{-1} \tilde{\boldsymbol{e}} \qquad \text{from Theorem 3.3.}
\end{aligned}$$

For a working variate associated with a non-residual parameter $\dot{\boldsymbol{H}}_i = \boldsymbol{Z}_i \dot{\boldsymbol{G}}_i \boldsymbol{Z}_i^T$ and

$$\begin{aligned}
\boldsymbol{q}_i &= \boldsymbol{Z}_i \dot{\boldsymbol{G}}_i \boldsymbol{Z}_i^T \boldsymbol{P}\boldsymbol{y} \\
&= \boldsymbol{Z}_i \dot{\boldsymbol{G}}_i \boldsymbol{G}_i^{-1} \boldsymbol{G}_i \boldsymbol{Z}_i^T \boldsymbol{P}\boldsymbol{y} \\
&= \boldsymbol{Z}_i \dot{\boldsymbol{G}}_i \boldsymbol{G}_i^{-1} \tilde{\boldsymbol{u}} \qquad \text{from Theorem 3.1.}
\end{aligned}$$

□

If $\boldsymbol{R}^{-1}$ and $\boldsymbol{G}^{-1}$ are of known form then these working variates reduce computation drastically.

3.6.2 *Alternative Forms for Restricted Likelihood and Scores*

So far attention has been given to the calculation of the descent directions. It was argued that the observed and expected information were computationally demanding and hence should be avoided in favour of the average information. However, the evaluation of the likelihood and score functions in (3.3.2), (3.4.1) and (3.4.2) are still computationally expensive. Fortunately alternative forms exist for these quantities which are less demanding.

Theorem 3.6. *The restricted log-likelihood in* (3.3.2) *can be expressed as*

$$\ell_r = -\frac{1}{2}\left((n-p)\log\sigma^2 + \log|\boldsymbol{G}| + \log|\boldsymbol{R}| + \log|\boldsymbol{C}| + \frac{1}{\sigma^2}\boldsymbol{y}^T\boldsymbol{P}\boldsymbol{y}\right)$$

where

$$\boldsymbol{C} = \begin{bmatrix} \boldsymbol{X}^T\boldsymbol{R}^{-1}\boldsymbol{X} & \boldsymbol{X}^T\boldsymbol{R}^{-1}\boldsymbol{Z} \\ \boldsymbol{Z}^T\boldsymbol{R}^{-1}\boldsymbol{X} & \boldsymbol{Z}^T\boldsymbol{R}^{-1}\boldsymbol{Z} + \boldsymbol{G}^{-1} \end{bmatrix}$$

and is the coefficient matrix in the mixed model equations (3.5.4)*.*

Proof. Consideration of only the determinants is necessary.

$$\begin{aligned}|\boldsymbol{C}| &= \left|\boldsymbol{Z}^T\boldsymbol{R}^{-1}\boldsymbol{Z}+\boldsymbol{G}^{-1}\right|\left|\boldsymbol{X}^T\boldsymbol{H}^{-1}\boldsymbol{X}\right| \\ &= \left|\boldsymbol{G}^{-1}\right|\left|\boldsymbol{I}_n-\boldsymbol{R}^{-1}\boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}^T\right|\left|\boldsymbol{X}^T\boldsymbol{H}^{-1}\boldsymbol{X}\right| && \text{using Result B.3} \\ &= \left|\boldsymbol{G}^{-1}\right|\left|\boldsymbol{R}^{-1}\right|\left|\boldsymbol{H}\right|\left|\boldsymbol{X}^T\boldsymbol{H}^{-1}\boldsymbol{X}\right|\end{aligned}$$

giving

$$\left|\boldsymbol{H}^{-1}\right|\left|\boldsymbol{X}^T\boldsymbol{H}^{-1}\boldsymbol{X}\right| = \left|\boldsymbol{G}^{-1}\right|\left|\boldsymbol{R}^{-1}\right|\left|\boldsymbol{C}\right|$$

□

Theorem 3.7. *(Cullis et al., 2006) The score equation for a residual variance parameter θ_i is*

$$U(\theta_i) = -\frac{1}{2}\left(\operatorname{tr}\left(\boldsymbol{R}^{-1}\dot{\boldsymbol{R}}_i\right) - \operatorname{tr}\left(\boldsymbol{C}^{-1}\boldsymbol{B}^T\boldsymbol{R}^{-1}\dot{\boldsymbol{R}}_i\boldsymbol{R}^{-1}\boldsymbol{B}\right) - \frac{1}{\sigma^2}\tilde{\boldsymbol{e}}^T\boldsymbol{R}^{-1}\dot{\boldsymbol{R}}_i\boldsymbol{R}^{-1}\tilde{\boldsymbol{e}}\right),$$

and the score equation for a non-residual variance parameter γ_i is

$$U(\gamma_i) = -\frac{1}{2}\left(\operatorname{tr}\left(\boldsymbol{G}_1^{-1}\dot{\boldsymbol{G}}_i\right) - \operatorname{tr}\left(\boldsymbol{G}_i^{-1}\dot{\boldsymbol{G}}_i\boldsymbol{G}_i^{-1}\boldsymbol{C}^{\boldsymbol{Z}_i\boldsymbol{Z}_i}\right) - \frac{1}{\sigma^2}\tilde{\boldsymbol{u}}^T\boldsymbol{G}_i^{-1}\dot{\boldsymbol{G}}_i\boldsymbol{G}_i^{-1}\tilde{\boldsymbol{u}}_i\right)$$

where $\boldsymbol{C}$ is the coefficient matrix in the mixed model equations (3.5.4)*, $\boldsymbol{B} = (\boldsymbol{X}, \boldsymbol{Z})$ and $\boldsymbol{C}^{\boldsymbol{Z}_i\boldsymbol{Z}_i}$ is the $r_i \times r_i$ partition of the inverse of $\boldsymbol{C}$ corresponding to $\boldsymbol{u}_i$.*

Proof. First consider $U(\theta_i)$. From (3.4.2) and using $\boldsymbol{P} = \boldsymbol{R}^{-1} - \boldsymbol{R}^{-1}\boldsymbol{B}\boldsymbol{C}^{-1}\boldsymbol{B}\boldsymbol{R}^{-1}$

$$\begin{aligned}U(\theta_i) &= -\frac{1}{2}\left(\operatorname{tr}\left((\boldsymbol{R}^{-1}-\boldsymbol{R}^{-1}\boldsymbol{B}\boldsymbol{C}^{-1}\boldsymbol{B}^T\boldsymbol{R}^{-1})\dot{\boldsymbol{R}}_i\right) - \frac{1}{\sigma^2}\boldsymbol{y}^T\boldsymbol{P}\boldsymbol{R}\boldsymbol{R}^{-1}\dot{\boldsymbol{R}}_i\boldsymbol{R}^{-1}\boldsymbol{R}\boldsymbol{P}\boldsymbol{y}\right) \\ &= -\frac{1}{2}\left(\operatorname{tr}\left(\boldsymbol{R}^{-1}\dot{\boldsymbol{R}}_i\right) - \operatorname{tr}\left(\boldsymbol{C}^{-1}\boldsymbol{B}^T\boldsymbol{R}^{-1}\dot{\boldsymbol{R}}_i\boldsymbol{R}^{-1}\boldsymbol{B}\right) - \frac{1}{\sigma^2}\tilde{\boldsymbol{e}}^T\boldsymbol{R}^{-1}\dot{\boldsymbol{R}}_i\boldsymbol{R}^{-1}\tilde{\boldsymbol{e}}\right) \\ & \qquad\qquad\qquad\qquad\qquad\qquad \text{from Theorem 3.3.}\end{aligned}$$

Now consider $U(\gamma_i)$. Define the $(p+r) \times r_i$ matrix $\boldsymbol{S}_i$ so that $\boldsymbol{B}\boldsymbol{S}_i = \boldsymbol{Z}_i$. The matrix $\boldsymbol{S}_i$ contains zeros everywhere except for an identity matrix of order r_i in the partition corresponding to $\boldsymbol{Z}_i$ in $\boldsymbol{B}$. Also define $\boldsymbol{G}^* = \boldsymbol{C} - \boldsymbol{B}^T\boldsymbol{R}^{-1}\boldsymbol{B}$. From (3.4.2) the trace term is

$$\begin{aligned}\operatorname{tr}\left(\boldsymbol{P}\dot{\boldsymbol{H}}_i\right) &= \operatorname{tr}\left(\dot{\boldsymbol{G}}_i\boldsymbol{Z}_i^T(\boldsymbol{R}^{-1}-\boldsymbol{R}^{-1}\boldsymbol{B}\boldsymbol{C}^{-1}\boldsymbol{B}^T\boldsymbol{R}^{-1})\boldsymbol{Z}_i\right) \\ &= \operatorname{tr}\left(\dot{\boldsymbol{G}}_i\boldsymbol{S}_i^T\boldsymbol{B}^T\boldsymbol{R}^{-1}(\boldsymbol{B}-\boldsymbol{B}\boldsymbol{C}^{-1}\boldsymbol{B}\boldsymbol{R}^{-1}\boldsymbol{B})\boldsymbol{S}_i\right) \\ &= \operatorname{tr}\left(\dot{\boldsymbol{G}}^{-1}\boldsymbol{S}_i^T\boldsymbol{B}^T\boldsymbol{R}^{-1}\boldsymbol{B}\boldsymbol{C}^{-1}(\boldsymbol{C}-\boldsymbol{B}\boldsymbol{R}^{-1}\boldsymbol{B})\boldsymbol{S}_i\right) \\ &= \operatorname{tr}\left(\dot{\boldsymbol{G}}_i\boldsymbol{S}_i^T\boldsymbol{B}^T\boldsymbol{R}^{-1}\boldsymbol{B}\boldsymbol{C}^{-1}\boldsymbol{G}^*\boldsymbol{S}_i\right) \\ &= \operatorname{tr}\left(\dot{\boldsymbol{G}}_i\boldsymbol{S}_i^T(\boldsymbol{C}-\boldsymbol{G}^*)\boldsymbol{C}^{-1}\boldsymbol{G}^*\boldsymbol{S}_i\right) \\ &= \operatorname{tr}\left(\dot{\boldsymbol{G}}_i\boldsymbol{S}_i^T\boldsymbol{G}^*\boldsymbol{S}_i\right) - \operatorname{tr}\left(\dot{\boldsymbol{G}}_i\boldsymbol{S}_i^T\boldsymbol{G}^*\boldsymbol{C}^{-1}\boldsymbol{G}^{-1}\boldsymbol{S}_i\right) \\ &= \operatorname{tr}\left(\dot{\boldsymbol{G}}_i\boldsymbol{G}_i^{-1}\right) - \operatorname{tr}\left(\boldsymbol{G}_i^{-1}\boldsymbol{C}^{\boldsymbol{Z}_i\boldsymbol{Z}_i}\boldsymbol{G}_i^{-1}\dot{\boldsymbol{G}}_i\right).\end{aligned}$$

The quadratic term may be written as

$$\begin{aligned}\boldsymbol{y}^T\boldsymbol{P}\dot{\boldsymbol{H}}_i\boldsymbol{P}\boldsymbol{y} &= \boldsymbol{y}^T\boldsymbol{P}\boldsymbol{Z}_i\dot{\boldsymbol{G}}_i\boldsymbol{Z}_i^T\boldsymbol{P}\boldsymbol{y}\\ &= \boldsymbol{y}^T\boldsymbol{P}\boldsymbol{Z}_i\boldsymbol{G}_i\boldsymbol{G}_i^{-1}\dot{\boldsymbol{G}}_i\boldsymbol{G}_i^{-1}\boldsymbol{G}_i\boldsymbol{Z}_i^T\boldsymbol{P}\boldsymbol{y}\\ &= \tilde{\boldsymbol{u}}_i^T\boldsymbol{G}_i^{-1}\dot{\boldsymbol{G}}_i\boldsymbol{G}_i^{-1}\tilde{\boldsymbol{u}}_i.\end{aligned}$$

Substitute these expressions into the score in (3.4.2) to obtain the result. □

Many of the terms required to calculate the restricted likelihood and the scores can be found during the process of absorption for the working variates, in particular $\log|\boldsymbol{C}|$ and $\boldsymbol{C}^{-1}$. This implies that the restricted log-likelihood and the score equations can be calculated with minimal computational effort. Computation can be further reduced by noticing that $\boldsymbol{C}$ and $\boldsymbol{C}^{-1}$ will typically contain a large number of zero elements. This sparsity can be exploited by using sparse matrix methods (Gilmour et al., 1995). This reduces computation further.

3.6.3 *The AI Algorithm*

The algorithm used to maximise the restricted log-likelihood (solve the score equations) is now succinctly stated.

1. Initiate variance parameters.

2. Solve mixed model equations in (3.5.4) for $\hat{\boldsymbol{\tau}}$, $\tilde{\boldsymbol{u}}$ and $\tilde{\boldsymbol{e}}$ given current estimates for variance parameters.

3. Form working variates using current estimates of variance parameters and predicted random effects via (3.5).

4. Augment mixed model equations with working variables and absorb to obtain elements of the average information matrix, $\log|\boldsymbol{C}|$ and $\boldsymbol{C}^{-1}$.

5. Calculate the score equations and the restricted log-likelihood in Theorems 3.7 and 3.6 respectively.

6. Update variance parameters using the descent direction defined by the current average information and score equations.

7. If restricted log-likelihood has not converged then go back to step 2 until convergence.

3.7 Analytical Approximations to the Marginal Likelihood

For models other than the standard mixed model, where both random effects and residuals are normally distributed, exact analytical forms for the marginal (restricted) likelihood are typically not available. In these cases approximations may be available and prove to be very useful. Here, two such approximations are presented. When used on the standard mixed model the first, called Laplace's method (Erdélyi, 1956 and de Bruijn, 1961), reproduces the full marginal likelihood. The second, called partial Laplace (Taylor, 2005 and Taylor & Verbyla, 2006), reproduces the restricted likelihood.

3.7.1 *Laplace's Method*

Consider the integral of the data over the r dimensional random effects vector $\boldsymbol{u}$ whose support is $\mathbb{R}^r$. Denote the exponent of the joint distribution as $m(\boldsymbol{y};\boldsymbol{\tau},\boldsymbol{u},\boldsymbol{\eta}) = m(\boldsymbol{u})$. For some distributions this will involve taking the exponent of the log-density. The marginal distribution is

$$f(\boldsymbol{y};\boldsymbol{\tau},\boldsymbol{\eta}) = \int_{\mathbb{R}^r} \exp\left(-m(\boldsymbol{u})\right) \partial\boldsymbol{u}.$$

The base idea behind Laplace's method is to approximate the integrand by a normal distribution, which has an integral of 1 when the area of integration is $\mathbb{R}^r$. In practice, this involves approximating the function $m(\boldsymbol{u})$ by a well chosen quadratic. The obvious choice for the quadratic is the second order Taylor polynomial around the maximum of the integrand $\tilde{\boldsymbol{u}}$. With this choice the first derivative of $m(\boldsymbol{u})$ is zero.

$$m(\boldsymbol{u}) \approx m(\tilde{\boldsymbol{u}}) - \frac{1}{2}(\boldsymbol{u}-\tilde{\boldsymbol{u}})^T \ddot{\boldsymbol{M}}(\tilde{\boldsymbol{u}})(\boldsymbol{u}-\tilde{\boldsymbol{u}})\partial\boldsymbol{u}$$

where $\ddot{\boldsymbol{M}}(\tilde{\boldsymbol{u}})$ the matrix of second derivatives evaluated at $\boldsymbol{u} = \tilde{\boldsymbol{u}}$. The approximate marginal integral is

$$\begin{aligned}
f(\boldsymbol{y};\boldsymbol{\tau},\boldsymbol{\eta}) &\approx \exp\left(-m(\tilde{\boldsymbol{u}})\right) \int_{\mathbb{R}^r} \exp\left(-\frac{1}{2}(\boldsymbol{u}-\tilde{\boldsymbol{u}})^T \ddot{\boldsymbol{M}}(\tilde{\boldsymbol{u}})(\boldsymbol{u}-\tilde{\boldsymbol{u}})\right)\partial\boldsymbol{u} \\
&= \exp\left(-m(\tilde{\boldsymbol{u}})\right)(2\pi)^{\frac{r}{2}} \left|\ddot{\boldsymbol{M}}(\tilde{\boldsymbol{u}})\right|^{-\frac{1}{2}} \times \\
&\qquad \int_{\mathbb{R}^r} \frac{1}{(2\pi)^{\frac{r}{2}}} \left|\ddot{\boldsymbol{M}}(\tilde{\boldsymbol{u}})\right|^{\frac{1}{2}} \exp\left(-\frac{1}{2}(\boldsymbol{u}-\tilde{\boldsymbol{u}})^T \ddot{\boldsymbol{M}}(\tilde{\boldsymbol{u}})(\boldsymbol{u}-\tilde{\boldsymbol{u}})\right)\partial\boldsymbol{u} \\
&= (2\pi)^{\frac{r}{2}} \left|\ddot{\boldsymbol{M}}(\tilde{\boldsymbol{u}})\right|^{-\frac{1}{2}} \exp\left(-m(\tilde{\boldsymbol{u}})\right).
\end{aligned}$$

This analytical approximation is free of the random effects, as they have been evaluated at their maximum. When this approximation is applied to the standard mixed model the full marginal likelihood is returned exactly. Indeed, for any density whose exponent is once or at most twice differentiable the approximation is exact.

3.7.2 *Partial Laplace's Method*

Laplace's method for analytical approximation, like all ML methods, will produce biased estimates for the variance parameters. Hence, the estimates for the fixed effects and the predictions for the random effects will also be biased. Laplace's method can be modified to reproduce the restricted likelihood for the standard mixed model. The modified version is called Partial Laplace (Taylor, 2005 and Taylor & Verbyla, 2006). It works by taking the expansion of the integrand around the maximum of the integrand free of fixed effects.

Partition the exponent of the joint distribution, using methods similar to those used to derive REML (Section 3.3), into two functions; one of which will not have the fixed effects $\boldsymbol{\tau}$ as an argument. That is

$$\begin{aligned} m(\boldsymbol{y};\boldsymbol{\tau},\boldsymbol{u},\boldsymbol{\eta}) &= m_1(\boldsymbol{y};\boldsymbol{\tau},\boldsymbol{u},\boldsymbol{\eta}) + m_2(\boldsymbol{y};\boldsymbol{u},\boldsymbol{\eta}) \\ &= m_1 + m_2 \qquad\qquad \text{say.} \end{aligned}$$

Expand the integrand of the marginal distribution around $\tilde{\boldsymbol{u}}_p$, the maximiser of m_2. The Taylor series expansion is then taken around this point. It is

$$m(\boldsymbol{u}) \approx m_1 + m_2 + (\boldsymbol{u} - \tilde{\boldsymbol{u}}_p)^T \dot{\boldsymbol{m}}_1 + \frac{1}{2}(\boldsymbol{u} - \tilde{\boldsymbol{u}}_p)^T \left(\ddot{\boldsymbol{M}}_1 + \ddot{\boldsymbol{M}}_2\right)(\boldsymbol{u} - \tilde{\boldsymbol{u}}_p)$$

where $\dot{\boldsymbol{m}}_1 = \dot{\boldsymbol{m}}_1(\boldsymbol{\tau}, \tilde{\boldsymbol{u}}_p)$ is the vector of derivatives of $m_1(\boldsymbol{\tau},\boldsymbol{u})$ evaluated at $\tilde{\boldsymbol{u}}_p$, $\ddot{\boldsymbol{M}}_1(\boldsymbol{\tau},\tilde{\boldsymbol{u}}_p)$ and $\ddot{\boldsymbol{M}}_2(\tilde{\boldsymbol{u}}_p)$ are the matrices of second derivatives of $m_1(\boldsymbol{\tau},\boldsymbol{u})$ and $m_2(\boldsymbol{u})$ respectively evaluated at $\tilde{\boldsymbol{u}}_p$.

Completing the square in the quadratic approximation of $m(\boldsymbol{y};\boldsymbol{\tau},\boldsymbol{u})$ using only the terms involving $\boldsymbol{u}$ gives

$$m(\boldsymbol{u}) \approx m_1 + m_2 + \frac{1}{2}\boldsymbol{a}^T \left(\ddot{\boldsymbol{M}}_1 + \ddot{\boldsymbol{M}}_2\right)\boldsymbol{a} - \frac{1}{2}\dot{\boldsymbol{m}}_1^T \left(\ddot{\boldsymbol{M}}_1 + \ddot{\boldsymbol{M}}_2\right)\dot{\boldsymbol{m}}_1$$

where $\boldsymbol{a} = (\boldsymbol{u} - \tilde{\boldsymbol{u}}_p) + \left(\ddot{\boldsymbol{M}}_1 + \ddot{\boldsymbol{M}}_2\right)^{-1}\dot{\boldsymbol{m}}_1$ and all instances of m_1 and m_2 and their derivatives are evaluated at $\tilde{\boldsymbol{u}}_p$. The approximate marginal density is

$$\begin{aligned} f(\boldsymbol{y};\boldsymbol{\tau},\boldsymbol{\eta}) &\approx \exp\left(-m_1 - m_2\right)\exp\left(\frac{1}{2}\dot{\boldsymbol{m}}_1^T \left(\ddot{\boldsymbol{M}}_1 + \ddot{\boldsymbol{M}}_2\right)^{-1}\dot{\boldsymbol{m}}_1\right) \times \\ &\quad \int_{\mathbb{R}^r} \exp\left(-\frac{1}{2}\boldsymbol{a}^T\left(\ddot{\boldsymbol{M}}_1 + \ddot{\boldsymbol{M}}_2\right)\boldsymbol{a}\right)\partial\boldsymbol{u} \\ &= (2\pi)^{\frac{r}{2}}\left|\ddot{\boldsymbol{M}}_1 + \ddot{\boldsymbol{M}}_2\right|^{-\frac{1}{2}}\exp\left(-m_1 - m_2\right)\exp\left(\frac{1}{2}\dot{\boldsymbol{m}}_1^T\left(\ddot{\boldsymbol{M}}_1 + \ddot{\boldsymbol{M}}_2\right)^{-1}\dot{\boldsymbol{m}}_1\right) \\ &= \left((2\pi)^r\left|\ddot{\boldsymbol{M}}_1\ddot{\boldsymbol{M}}_2^{-1} + \boldsymbol{I}_p\right|^{-\frac{1}{2}}\exp\left(-m_1(\tilde{\boldsymbol{u}}) + \frac{1}{2}\dot{\boldsymbol{m}}_1(\ddot{\boldsymbol{M}}_1 + \ddot{\boldsymbol{M}}_2)^{-1}\dot{\boldsymbol{m}}_1\right)\right) \times \\ &\qquad\qquad \left(\left|\ddot{\boldsymbol{M}}_2\right|^{-\frac{1}{2}}\exp\left(-m_2(\tilde{\boldsymbol{u}}_p)\right)\right) \\ &= m_1^*(\boldsymbol{y};\boldsymbol{\tau},\tilde{\boldsymbol{u}}_p)m_2^*(\boldsymbol{y};\tilde{\boldsymbol{u}}_p) \qquad\qquad \text{say.} \end{aligned}$$

The functions $m_1^*(\boldsymbol{y};\boldsymbol{\tau},\tilde{\boldsymbol{u}}_p)$ and $m_2^*(\boldsymbol{y};\tilde{\boldsymbol{u}}_p)$ are analogous to the conditional and marginal likelihoods derived for restricted likelihood. That is $m_2^*(\boldsymbol{y};\tilde{\boldsymbol{u}}_p)$ is a function of the data free of fixed effects used to estimate the variance parameters. The function $m_1^*(\boldsymbol{y};\boldsymbol{\tau},\tilde{\boldsymbol{u}}_p)$ is then used to estimate the fixed effects given the data free of fixed effects.

In the standard mixed model $\tilde{\boldsymbol{u}}_p$ is a maximiser of $m_1(\boldsymbol{y}; \boldsymbol{\tau}, \boldsymbol{u})$ as well as $m_2(\boldsymbol{y}; \boldsymbol{u})$ and the approximation $m_2(\boldsymbol{y}; \tilde{\boldsymbol{u}}_p)$ reproduces the restricted likelihood in (3.3.2) exactly (Taylor, 2005 and Taylor & Verbyla, 2006). In general it is expected that if $\tilde{\boldsymbol{u}}_p$ maximises both partitions of the data then good results will be obtained. Further, if the exponent of the integrand is only once or twice differentiable then the Taylor series expansion is exact.

The resulting function for estimating the dispersion parameters obtained using the partial Laplace method has an alternative derivation. It is exactly the approximation obtained by partitioning the observed outcomes, ignoring the part of the density corresponding to m_1 and then integrating the remaining function.

Chapter 4

Review: Subset Selection and Biased Estimation

4.1 Introduction

Aspects of subset selection and biased estimation in linear models are discussed in this chapter. These methods have benefits for both prediction of new observations and model interpretation. Also, they have direct application to QTL identification when multiple QTL models are used. When the parameter estimates are required to be unbiased it will be shown that it may be important to use a reduced set of explanatory variables. This decreases the variance of the predictions (Miller, 2002) and makes interpretation of the model as parsimonious as possible. Various methods for reducing the number of explanatory variables in the model, such as the familiar forward and backward selection, will be discussed. If the requirement of unbiased estimates is relaxed then other methods of estimating the linear model's parameters can be useful. Two of these approaches are ridge regression (Hoerl & Kennard, 1970a and Hoerl & Kennard, 1970b) and the LASSO (Tibshirani, 1996). Ridge regression is also commonly used in situations where the correlations between the explanatory variables are high. In such cases the ridge estimates have substantially lower variance than their unbiased counterparts. Both ridge regression and the LASSO are commonly considered as constrained regression problems, however it will be shown that both these methods can also be specified as a penalised regression and a random effects model where the random effects follow a normal and a double exponential distribution respectively. Estimation of LASSO effects requires non-standard methods as the objective function is non-differentiable. The interior point descent algorithm of Osborne et al. (2000) is used throughout this thesis and will be reviewed in this chapter.

4.2 Why Select Subsets?

Consider the linear model

$$\boldsymbol{y} = \boldsymbol{L}\boldsymbol{\beta} + \boldsymbol{e} \tag{4.2.1}$$

where $\boldsymbol{y}$ is the $n \times 1$ vector of outcomes, $\boldsymbol{L}$ is the $n \times q$ matrix of explanatory variables, $\boldsymbol{\beta}$ is the $q \times 1$ vector of location parameters and $\boldsymbol{e}$ is the $n \times 1$ vector of residuals. Commonly, the residuals are assumed to be independently and identically distributed. There are two, not necessarily exclusive, uses for the linear model: prediction of future observations and interpretation of the biological system.

For the present discussion, assume that the location parameter vector $\boldsymbol{\beta}$ is estimated by the least squares method. This gives unbiased estimates, their expected values and their variance as

$$\begin{aligned} \hat{\boldsymbol{\beta}} &= (\boldsymbol{L}^T\boldsymbol{L})^{-1}\boldsymbol{L}^T\boldsymbol{y}, \\ \mathrm{E}\left(\hat{\boldsymbol{\beta}}\right) &= \boldsymbol{\beta} \qquad\qquad \text{and} \\ \mathrm{var}\left(\hat{\boldsymbol{\beta}}\right) &= \sigma^2(\boldsymbol{L}^T\boldsymbol{L})^{-1} \end{aligned} \tag{4.2.2}$$

where σ^2 is the residual variance. The residual variance is typically unknown and must be estimated. The common (unbiased) estimate is

$$\hat{\sigma}^2 = \frac{(\boldsymbol{y}-\boldsymbol{L}\hat{\boldsymbol{\beta}})^T(\boldsymbol{y}-\boldsymbol{L}\hat{\boldsymbol{\beta}})}{n-q} = \frac{\boldsymbol{y}^T\left(\boldsymbol{I}-\boldsymbol{L}(\boldsymbol{L}^T\boldsymbol{L})^{-1}\boldsymbol{L}^T\right)\boldsymbol{y}}{n-q}.$$

This is the REML estimate and not the ML estimate described in Chapter 3.

4.2.1 *Prediction*

The linear model in (4.2.1), estimated by least squares, can be used for predicting a new observation from a previously fitted model. Intuitively, it could be argued that all of the available explanatory variables should be included in the model, as all the variables may have some effect on the outcome. However, this may not be the best strategy, as a prediction from a model with a larger number of explanatory variables will have a higher variance than a prediction from a model with a smaller set of explanatory variables.

Theorem 4.1. *The variance of a prediction is greater from model A than from model A^* where the explanatory variables in model A^* are a subset of the explanatory variables in model A.*

Proof. (Adapted from Miller, 2002). For a given observation y_i with explanatory variables $\boldsymbol{l}_i$ and least squares estimate $\hat{\boldsymbol{\beta}}$, the prediction $\hat{y}_i$ and the prediction's variance from the fitted model are

$$\begin{aligned} \hat{y}_i &= \boldsymbol{l}_i^T\hat{\boldsymbol{\beta}} \\ \mathrm{var}\left(\hat{y}_i\right) &= \sigma^2\boldsymbol{l}_i^T(\boldsymbol{L}^T\boldsymbol{L})^{-1}\boldsymbol{l}_i. \end{aligned}$$

Consider a Cholesky decomposition of the matrix $\boldsymbol{L}^T\boldsymbol{L}$ into $\boldsymbol{K}\boldsymbol{K}^T$ where $\boldsymbol{K}$ is a square lower triangular matrix of size q. Note that with this decomposition $(\boldsymbol{L}^T\boldsymbol{L})^{-1} = (\boldsymbol{K}^{-1})^T\boldsymbol{K}^{-1}$. This gives the variance of the prediction to be $\sigma^2(\boldsymbol{K}^{-1}\boldsymbol{l}_i)^T(\boldsymbol{K}^{-1}\boldsymbol{l})$.

Write $\boldsymbol{L} = [\boldsymbol{L}_{A^*}\boldsymbol{L}_{\bar{A}^*}]$ so that $\boldsymbol{L}_{A^*}$ is the design matrix for the variables in model A^* and $\boldsymbol{L}_{\bar{A}^*}$ is the design matrix for all the explanatory variables not in model A^*. Partition $\boldsymbol{K}$, $\boldsymbol{K}^{-1}$, $\boldsymbol{l}_i$ and $\hat{\boldsymbol{\beta}}$ comformably using the sets A^* and $\bar{A}^*$. The variance of the prediction from model A^* is

$$\operatorname{var}\left(\boldsymbol{l}_{iA^*}^T\hat{\boldsymbol{\beta}}_{A^*}\right) = \sigma^2\left(\boldsymbol{K}_{A^*}^{-1}\boldsymbol{l}_{iA^*}\right)^T\left(\boldsymbol{K}_{A^*}^{-1}\boldsymbol{l}_{iA^*}\right),$$

which is the sums of squares of the q_{A^*} explanatory variables in $\boldsymbol{L}_{A^*}$. Hence, the prediction's variance is a monotone increasing function of the number of explanatory variables included in the model and $\operatorname{var}\left(\boldsymbol{l}_{iA^*}^T\hat{\boldsymbol{\beta}}_{A^*}\right) \leq \operatorname{var}\left(\boldsymbol{l}_{iA}^T\hat{\boldsymbol{\beta}}_{A}\right)$. □

The result from Theorem 4.1 implies that models containing very few explanatory variables are favourable as they have low prediction variance. However, overly small models may have inflated bias in both the parameter estimates and the predicted values.

Theorem 4.2. *Suppose that the expected value of the outcomes is* $\mathrm{E}(\boldsymbol{y}) = \boldsymbol{L}\boldsymbol{\beta}$. *A model containing a subset of the explanatory variables will have a biased least squares estimate, unless all the effects of variables not included in the model are identically zero. Further, the predictions from such an overly reduced model will also be biased unless all effects not included in the model are identically zero.*

Proof. (Adapted from Miller, 2002). Let the explanatory variables included in the reduced model be arranged in the design matrix by $\boldsymbol{L}_A$ and the one not included in the reduced model be denoted by $\boldsymbol{L}_{\bar{A}}$. Arrange the explanatory variables and the model's effects so that $\boldsymbol{L} = [\boldsymbol{L}_A, \boldsymbol{L}_{\bar{A}}]$ and $\boldsymbol{\beta} = [\boldsymbol{\beta}_A^T\boldsymbol{\beta}_{\bar{A}}^T]^T$. The expected value of the reduced model's estimates is

$$\begin{aligned}
\mathrm{E}\left(\hat{\boldsymbol{\beta}}_A\right) &= (\boldsymbol{L}_A^T\boldsymbol{L}_A)^{-1}\boldsymbol{L}_A^T\mathrm{E}(\boldsymbol{y}) \\
&= (\boldsymbol{L}_A^T\boldsymbol{L}_A)^{-1}\boldsymbol{L}_A^T\boldsymbol{L}\boldsymbol{y} \\
&= (\boldsymbol{L}_A^T\boldsymbol{L}_A)^{-1}\boldsymbol{L}_A^T[\boldsymbol{L}_A\boldsymbol{L}_{\bar{A}}]\boldsymbol{\beta} \\
&= (\boldsymbol{L}_A^T\boldsymbol{L}_A)^{-1}\boldsymbol{L}_A^T\boldsymbol{L}_A\boldsymbol{\beta}_A + (\boldsymbol{L}_A^T\boldsymbol{L}_A)^{-1}\boldsymbol{L}_A^T\boldsymbol{L}_{\bar{A}}\boldsymbol{\beta}_{\bar{A}} \\
&= \boldsymbol{\beta}_A + (\boldsymbol{L}_A^T\boldsymbol{L}_A)^{-1}\boldsymbol{L}_A^T\boldsymbol{L}_{\bar{A}}\boldsymbol{\beta}_{\bar{A}}.
\end{aligned}$$

So the least squares estimate $\hat{\boldsymbol{\beta}}_A$ is biased unless $\boldsymbol{\beta}_{\bar{A}} = \boldsymbol{0}$.

For a given observation y_i and associated explanatory variables $\boldsymbol{l}_i^T = [\boldsymbol{l}_{iA}^T, \boldsymbol{l}_{i\bar{A}}^T]$ the bias in the prediction is

$$\begin{aligned}
\boldsymbol{l}_i\boldsymbol{\beta} - \mathrm{E}\left(\boldsymbol{l}_{iA}^T\hat{\boldsymbol{\beta}}_{iA}\right) &= \boldsymbol{l}_i\boldsymbol{\beta} - \boldsymbol{l}_{iA}^T\boldsymbol{\beta}_A - \boldsymbol{l}_{iA}^T(\boldsymbol{L}_A^T\boldsymbol{L}_A)^{-1}\boldsymbol{L}_A^T\boldsymbol{L}_{\bar{A}}\boldsymbol{\beta}_{\bar{A}} \\
&= \boldsymbol{l}_{i\bar{A}}^T\boldsymbol{\beta}_{\bar{A}} - \boldsymbol{l}_{iA}^T(\boldsymbol{L}_A^T\boldsymbol{L}_A)^{-1}\boldsymbol{L}_A^T\boldsymbol{L}_{\bar{A}}\boldsymbol{\beta}_{\bar{A}}.
\end{aligned}$$

The predictions from a reduced model fitted by least squares will have biased predictions unless $\boldsymbol{\beta}_{\bar{A}} = \boldsymbol{0}$. □

Theorem 4.1 indicates that a small model is beneficial for reducing the prediction's variance and Theorem 4.2 indicates that a large model is beneficial for reducing the prediction's bias. The goal of model selection is to minimise the prediction variance while not removing too many variables to make the predictions overly biased.

4.2.2 *Interpretation of Model*

Now consider the use of linear models, estimated by least squares, for drawing a useful and plausible explanation of the biological system under consideration. Obviously, the model should only contain explanatory variables that have an effect on the outcome, all other explanatory variables are not interesting. Further, the estimated effects included in the model should represent their unknown counterparts. Theorem 4.2 shows that an estimate of a location parameter is biased if important explanatory variables are removed from the model. Hence, the aim of model selection for explanation is to find the simplest model whose effect estimates are not overly biased.

4.3 Algorithms for Finding Reduced Linear Models

In the previous section it was shown that model selection is vital for both predictive models and models for interpretation. There have been many methods proposed to find the *best* model or models. Only the most popular are reviewed here. Much of this section is taken from Seber (1977), Neter et al. (1996) and Miller (2002) where the reader will find a more detailed discussion on these and other algorithms.

The algorithms can be broken up into three broad categories. The first are the forward moving algorithms that start with a model containing only a mean and add variables until a final model is found. The next is the backwards moving algorithm that starts with a full model (all the explanatory variables included in the model) and removes explanatory variables until a final model is reached. The last are the search routines that systematically assess all models containing a given number of variables.

For explanatory models, and to a lesser extent predictive models, there are a number of restrictions which the analyst should observe while building the reduced model, irrespective of the type of algorithm used. The principle of marginality and functional marginality should be respected (Nelder, 1994). The principle of marginality stipulates that interactions should not be included in the model unless the main effects are included. Functional marginality stipulates, for example, that a quadratic term should not be fitted unless the corresponding linear term is also fitted. Both these types of marginality circumvent unwanted restrictions on the model parameter's solution space. Also, individual dummy variables for a factor should not be included in the model unless all of the dummy variables for that factor are included.

4.3.1 *Forward Selection*

This is a forward moving algorithm, that starts with a null model (containing only a mean parameter) and sequentially adds variables until the final model is found. An arbitrary critical F-value, called F-to-enter or F_e, is specified prior to the commencement of the algorithm. The first step is to assess all the variables individually by adding them to the null model.

That is, fitting the set of models

$$\boldsymbol{y} = \mu + \boldsymbol{l}_j\beta_j + \boldsymbol{e}; \qquad\qquad j = 1 \dots q.$$

The set of F-ratios corresponding to these q models are calculated

$$F_j^* = \frac{MSR_j}{MSE_j}$$

where MSR_j is the mean squares due to the j^{th} model and MSE_j is the mean squared error from the j^{th} model.

The variable that has the highest F_j^* value is added to the model if it exceeds F_e. The process is repeated by adding the remaining explanatory variables to the previously enlarged model. The only difference is that the F_j^* values are now the partial F-ratios: the mean squared residual and the mean squared error for the j^{th} explanatory variable conditional on the other explanatory variables being in the model.

The algorithm terminates when all explanatory variable that are not included in the model have F^* values that do not exceed F_e. Convergence is guaranteed as there are a finite number of variables to add into the model.

If the explanatory variables are a mix of both covariates and factors then the critical F_e should be replaced with a p-value to enable all variables to be compared meaningfully. If some of the variables are marginal or functionally marginal to others, then they should not be included into the model before the marginal variables. This implies that the set of potential explanatory variables in such a case will grow with the inclusion of a variable.

A word of caution is given regarding the meaning of the statistics $\{F_j^*\}_{j=1}^q$ or the corresponding p-values. It is commonly assumed that these follow an F-distribution under the null hypothesis of no-effect for variable $\boldsymbol{l}_j$. An individual F_j^* statistic for a variable $\boldsymbol{l}_j$ chosen at random from all the potential explanatory variables will follow an F-distribution. Their maximum, the F_j^* value corresponding to the variable most associated with the outcome, will not have an F-distribution and hence any test assuming this will be incorrect (Draper et al., 1971; Seber, 1977; and Miller, 2002).

A further problem with any subset selection method is estimating the residual variance for use in the significance test. Early in a forward moving selection algorithm the residual variance estimate may be artificially inflated. Late in the forward moving selection procedure the residual variance estimate may be underestimated.

4.3.2 *Backward Selection*

This is a backwards moving algorithm. It starts with a maximal, or full, model containing all the explanatory variables considered for inclusion into the final regression model. Unlike forward selection, backward selection removes explanatory variables from the model. Prior to the commencement of the algorithm an arbitrary constant is chosen. This is often called F-to-remove or F_r.

The first step is to estimate a model with all the explanatory variables present. The next step is to calculate the partial F-statistics $\{F_j^\dagger\}_{j=1}^q$ formed by dropping each explanatory variable from the model in turn. The explanatory variable corresponding to the smallest $F_j^\dagger$ is removed if $F_j^\dagger$ is smaller than F_r.

The process is repeated until all explanatory variables in the model have partial F-statistics greater than F_r. Convergence is guaranteed as there are only a finite number of variables to remove from the model.

Much of the discussion of the forward selection algorithm applies to the backward selection algorithm. In particular: if the explanatory variables are a mix of covariates and factors then the corresponding partial p-value should be used instead of $\{F_j^\dagger\}_{j=1}^q$. The true significance level is something different from that given from $F_j^\dagger$.

Backward selection is more computationally expensive than forward selection as the cost of fitting each model is greater, especially if the number of explanatory variables is large. If the number of explanatory variables is greater than the number of observations the least squares estimates are not uniquely defined and hence the algorithm is not appropriate.

4.3.3 *Stepwise Selection*

Forward and backward selection can be seen as simplifications of a slightly more general algorithm, namely stepwise regression (Efroymson, 1960). This algorithm is generally seen as a forward moving algorithm but it does include some backwards steps. Like forward selection and backward selection F_e and F_r are specified prior to the commencement of the algorithm.

The algorithm starts by finding the explanatory variable, if any, that is most significantly associated with the outcome. That is, finding the F_j^* that exceeds F_e by the most. After this process has been repeated twice, so that there are two variables in the model, each of the two variables are assessed for removal by calculating $F_j^\dagger$. If either or both variables have a smaller $F_j^\dagger$ than F_r then the smallest is removed.

The algorithm continues in this fashion, adding a new explanatory variable and then assessing previously chosen explanatory variables, until no explanatory variable absent from the model has a F_j^* value exceeding F_e and no explanatory variable present in the model has an $F_j^\dagger$ smaller than F_r. Convergence is guaranteed if $F_r < F_e$ (Miller, 2002).

As observed for forward and backward selection, special consideration should be given when the explanatory variables are a mix of covariates and factors. Also, like forward and backward selection, the F-statistics will have an unknown distribution and hence the F-tests at each stage of the model building process will be incorrect (Draper et al., 1971; Seber, 1977; and Miller, 2002).

4.3.4 *Best Subsets*

Forward selection, backward selection and stepwise will often produce different models. Further, there is no guarantee that any of these models will produce the best model. This is because none of these algorithms search the entire model space.

This suggests the use of methods that search the model space more thoroughly. The complete search of the model space is a method called best subsets (or all subsets). Methods to identify all the models and compute the least squares estimates efficiently are given in Seber (1977) and Miller (2002). Fitting all the subsets is a computationally demanding task, much more demanding than any of the previous model identification methods. However, some methods such as Branch and Bound (e.g. Miller, 2002) try to efficiently search the parameter space by only considering those models which are likely to be good. This is a variation on the same theme generating the selection and stepwise algorithms.

Once all the subset models have been identified and estimated they need to be individually assessed for *goodness*. Generally goodness is defined by the model's ability to predict the data. Sometimes the number of parameters in the model are taken into account and the models that can predict the data well, with fewer parameters, are favoured. There are three common statistics that use all the available data for estimation for assessing the goodness of a model: the coefficient of determination (R_p^2); the adjusted coefficient of determination (Adj-R_p^2); and Mallows' C_p statistic. A brief description of these will be given. More detail can be found in Seber (1977) and Neter et al. (1996).

The methods, based on the coefficients of determination statistics, simply try to find the models that predict the data the best. The two coefficients of determination statistics for a model with p parameters are

$$\begin{aligned} R_p^2 &= \frac{SSR}{SST} = 1 - \frac{SSE}{SST} \\ \text{Adj-}R_p^2 &= 1 - \frac{(n-1)}{(n-p)}\frac{SSE}{SST} \end{aligned}$$

where SSR is the sums of squares explained by the model, SSE is the residual sums of squares and SST is the total sums of squares. The Adj-R_p^2 statistic penalises those models with greater numbers of parameters. However, the penalty is minimal when the number of parameters is much smaller than the number of observations.

Adding more explanatory variables into a model will increase the R_p^2 and Adj-R_p^2 statistics, assuming the adjustment for Adj-R_p^2 is not great. Hence the best subset, using either of these statistics, is the set where the addition of explanatory variables does not increase R_p^2 and/or Adj-R_p^2 dramatically.

An alternative is Mallows' C_p statistic, defined as

$$C_p = \frac{SSE_p}{MSE_q} - (n - 2p)$$

where SSE_p is the residual sums of squares for the model which has size p and MSE_q is the mean squared error for the model containing all the explanatory variables. The denominator is chosen to be an unbiased estimator of the residual variance.

It can be shown (Neter et al., 1996) that when there is no bias in the predictions - that is, $\mathrm{E}\left(\hat{y}_i\right) = \mu_i$ for $i = 1 \ldots n$ and μ_i is the true observation without error, the expected value of C_p is approximately p. A model with a value of C_p far from p is likely to have biased predictions and a model with C_p close to p is likely to produce good predictions.

4.3.5 *One True Model?*

One appealing aspect of forward selection, backwards selection and stepwise selection is that these algorithms will produce a single model for interpretation and prediction. On the other hand the best subsets method may identify multiple models which fit the data approximately equivalently. Hence, there may be no single model that describes the data better than other models. This can be more realistic from a data analysis point of view.

4.4 Ill-Conditioning

In many linear models the explanatory variables are correlated, sometimes highly correlated. If these highly correlated variables are included simultaneously in the model then the system of equations used for estimation may become ill-conditioned. This has the effect of making the estimates highly unstable, where small changes in the data produce potentially large changes in estimation. The estimates and the predictions may have unreasonably high variance.

Ill-conditioning arises because the explanatory variables in the design matrix are highly correlated. In extreme situations at least one of the explanatory variables is an approximate linear combination of other explanatory variables. Consider the total variance, the sum of the marginal variances, of the p fixed effects' estimates given in (4.2.2)

$$\begin{aligned}\sum_{i=1}^{q} \mathrm{var}\left(\hat{\beta}_i\right) &= \sigma^2 \mathrm{tr}\left(\left(\boldsymbol{L}^T\boldsymbol{L}\right)^{-1}\right)\\ &= \sigma^2 \sum_{i=1}^{q} \lambda_i^{-1}\\ &\geq \sigma^2 \lambda_s^{-1}\end{aligned}$$

where $\{\lambda_i\}_{i=1}^q$ is the set of eigenvalues for $(\boldsymbol{L}^T\boldsymbol{L})^{-1}$ and λ_s is the smallest of this set (Seber, 1977). Since $(\boldsymbol{L}^T\boldsymbol{L})^{-1}$ is positive definite then all the eigenvalues will be positive. This variance may be unreasonably large if the smallest eigenvalue is very small.

Many methods are used to remove the unwanted outcomes of ill-conditioning. These include: model reduction, prudently choosing explanatory variables which are not ill-conditioned; data transformation, reducing the explanatory variables to principal components which are in turn used as explanatory variables; and biased estimation, relaxing the unbiasedness requirement of the estimators. This last method will be discussed in more detail in the next section. Biased estimation works because the ill-conditioned system is transformed into a non ill-conditioned system. Relaxing the unbiasedness of the estimators requirement is often desirable, especially for prediction purposes.

4.5 Biased Estimation

One of the key attributes of the fixed effect estimates, such as least squares, is that they are unbiased in the sense that $\mathrm{E}\left(\hat{\boldsymbol{\beta}}\right) = \boldsymbol{\beta}$. However, if this constraint is removed the estimates obtained from an ill-conditioned system can be more stable. The estimates themselves produce predictions with lower mean squared error (MSE)

$$\begin{aligned}
\mathrm{MSE} &= \mathrm{E}\left((y_i - \mu_i)^2\right) \\
&= \mathrm{E}\left((y_i - \mathrm{E}\,(y_i) + \mathrm{E}\,(y_i) - \mu_i)^2\right) \\
&= \mathrm{E}\left((y_i - \mathrm{E}\,(y_i))^2\right) + 2B\mathrm{E}\,(y_i - \mathrm{E}\,(y_i)) + B^2 \\
&= \mathrm{E}\left((y_i - \mathrm{E}\,(y_i))^2\right) + B^2 \\
&= \sigma^2 + B^2
\end{aligned}$$

where σ^2 is the residual variance and B is the bias - the distance from the expected value to the true value (Neter et al., 1996). Unbiased estimates require that $\mathrm{E}\,(y_i) = \mu_i$ or equivalently $B = 0$. This is a constraint when considering estimation for minimum MSE. If this constraint is lifted then the estimates will potentially have lower MSE. This attribute makes biased estimates more appealing than unbiased ones for prediction purposes and for situations when the estimate's variance is unwarrantedly large, as it is from an ill-conditioned system.

There are three common methods used for biased estimation: constrained regression; penalised regression; and random effect models. There is an equivalence between all three methods given certain assumptions and choices for constraining parameters. The different methods of estimation are essentially different ways to specify the ratio of the variance parameters of the random effects model.

The most common type of biased estimates are the ridge regression estimates (Hoerl & Kennard, 1970a and Hoerl & Kennard, 1970b). Another type of biased estimate is the least absolute selection and shrinkage operator (LASSO; Tibshirani, 1996). Both these types of biased estimation can be defined by all three mentioned methods of biased estimation. In fact, they can both be seen to be special instances of another class of estimators called bridge estimators (Frank & Friedman, 1993 and Fu, 1998).

Both ridge regression and LASSO estimates have been proven to reduce MSE for sensible choices of the shrinkage parameters (Hoerl & Kennard, 1970b; Huang, 2003; and Rosset & Zhu, 2004). This property makes these methods popular.

The three methods of biased estimation will now be discussed using ridge regression and the LASSO as examples. These models are of particular interest as they have the favourable quality of simplicity. For illustration assume the linear model in (4.2.1). Further, assume that all the explanatory variables in the design matrix $\boldsymbol{L}$ have equal variance or that they can be sensibly transformed to have equal variance. This is important as the estimation methods are not invariant to arbitrary changes of scale.

4.5.1 *Constrained Likelihood*

A seemingly arbitrary method for producing biased estimates is constrained likelihood. This approach involves minimising an objective function, such as the sums of squares, subject to a constraint on the parameters $\boldsymbol{\beta}$. When the objective function is the sums of squares the method is called constrained regression. Typically the constraint is formulated by limiting the length of the parameters. For ridge regression length is defined as the L_2-norm, for the LASSO it is defined as the L_1-norm and for the bridge it is the L_p-norm (see Definition C.1). The constrained estimation problem for a q dimensional vector $\boldsymbol{\beta}$ can be stated as

$$\begin{array}{ll} \underset{\boldsymbol{\beta} \in \mathbb{R}^q}{\text{minimise:}} & S(\boldsymbol{\beta}) = \frac{1}{2}(\boldsymbol{y} - \boldsymbol{L}\boldsymbol{\beta})^T(\boldsymbol{y} - \boldsymbol{L}\boldsymbol{\beta}) \\ \text{subject to:} & \|\boldsymbol{\beta}\|_p \leq t \end{array}$$

where $0 \leq t < t_o$, $t_o = \|\hat{\boldsymbol{\beta}}\|_p$, $\hat{\boldsymbol{\beta}}$ is the least squares estimate of $\boldsymbol{\beta}$ and $\|\cdot\|_p$ is the L_p-norm operator.

The constrained estimation process can be displayed graphically when there are $q = 2$ explanatory variables. For ridge regression the solution space is the L_2 circle centred at the origin and for the LASSO it is the L_1 circle centred at the origin. In both cases the solution is the point where the ellipses defining the objective function for different choices of $\boldsymbol{\beta}$ intersect the solution space (Tibshirani, 1996). This is shown in Figure 4.1.

After viewing Figure 4.1 it becomes immediately obvious that the nature of the estimates from ridge regression and the LASSO have similarities. However, it is also clear that the estimates from the LASSO method may contain identically zero elements, induced by the polyhedral nature of the L_1 solution space. In the example variable 1 is estimated to be zero, while variable 2 is non-zero and shrunk towards zero. The ridge does not have this ability to produce identically zero estimates, as the L_2 solution space is not polyhedral.

The choice of the constraint parameter t can be quite arbitrary. Commonly, cross validation or some variant of it is used to choose this parameter. This method chooses t such that it predicts the data best among all choices of t.

A constraint parameter for ridge regression that has been suggested as useful in nearly all cases is (Hoerl et al., 1975)

$$t = \frac{q\sigma^2}{\boldsymbol{\beta}^T\boldsymbol{\beta}}.$$

The values σ^2 and $\boldsymbol{\beta}$ are unknown and are replaced by their least squares estimates, if they are available. This can be misleading as often the estimates will have high variance and hence may be far from the true values.

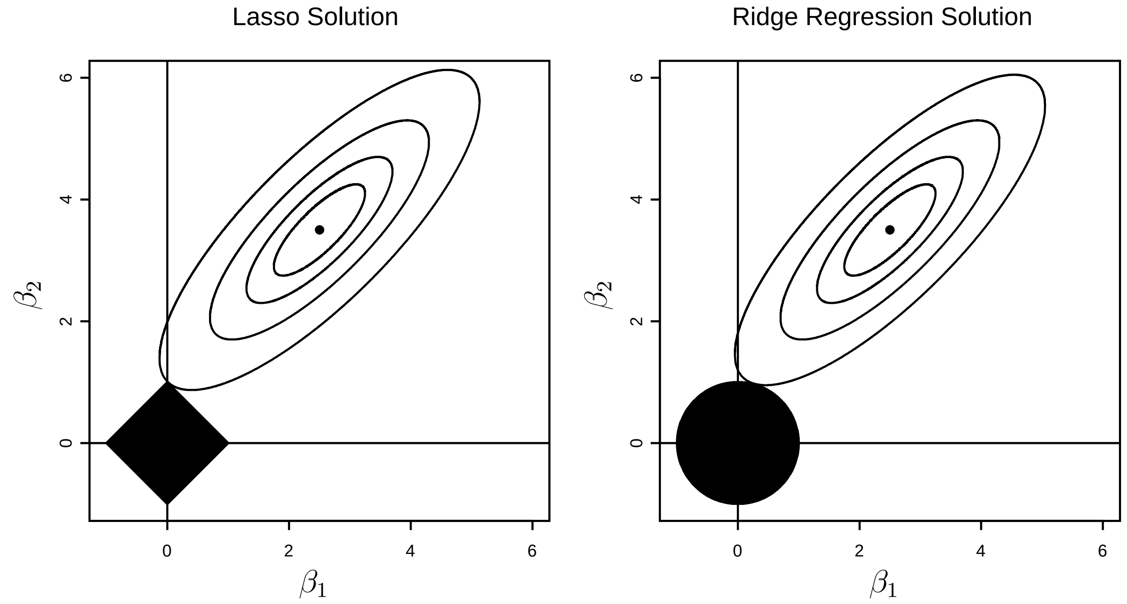

Figure 4.1: LASSO and Ridge Regression solution spaces with objective function ellipses. Ellipses are centred around the unconstrained estimates. Constrained estimates are the points where the ellipses first reach the solution space.

As the name suggests the LASSO performs selection as well as shrinkage. As such, it may replace selection methods in some situations and is in direct competition with other automatic model selection and estimation methods such as the non-negative garotte (Breiman, 1995) and the elastic net (Zou & Hastie, 2005).

4.5.2 *Penalised Likelihood*

The constraint problem can be viewed as a quadratic or more generally a convex programming problem. The Lagrangian function plays a key role in methods for constrained estimation. It incorporates the constraint into the objective function, giving a single function to optimise. For the constrained regression problem, ignoring those terms not involving $\boldsymbol{\beta}$, the Lagrangian is

$$\mathscr{L}(\boldsymbol{\beta}, \lambda) = S(\boldsymbol{\beta}) + \lambda \|\boldsymbol{\beta}\|_p \tag{4.5.1}$$

where λ is the Lagrange multiplier.

The function (4.5.1) is also called a penalised regression if $S(\boldsymbol{\beta})$ is the sums of squares objective. If $S(\boldsymbol{\beta})$ is a log-likelihood then it is called a penalised likelihood. The estimates are taken to be those that minimise the function $\mathscr{L}(\boldsymbol{\beta}, \lambda)$ for a given penalty parameter λ. There is a one to one relationship between the penalty parameter λ and the constraint parameter t. If there was no such relationship then the Lagrangian would not be a unique description of the constrained estimation process.

Unfortunately the exact functional relationship between the constraint parameter t and the penalty parameter λ is unknown for the LASSO. If an equivalent t for a given λ, or vice versa, is required then a numerical search is required. Osborne et al. (2000) suggest using a grid search or a Newton-Raphson iterative method to solve this problem. The latter is used in the implementation of the computational methods for this thesis; details are given in Chapter 6.

4.5.3 *Random Effects Models*

Consider a model where the effects are identically and independently normally distributed with mean zero and variance σ_g^2. In this case the log-joint distribution of the outcomes and effects, ignoring terms not involving $\boldsymbol{\beta}$, is

$$\begin{aligned}\log f(\boldsymbol{y}, \boldsymbol{\beta}) &= -\frac{1}{2\sigma^2}(\boldsymbol{y} - \boldsymbol{L\beta})^T(\boldsymbol{y} - \boldsymbol{L\beta}) - \frac{1}{2\sigma_g^2}\boldsymbol{\beta}^T\boldsymbol{\beta} \\ &\propto (\boldsymbol{y} - \boldsymbol{L\beta})^T(\boldsymbol{y} - \boldsymbol{L\beta}) + \frac{\sigma^2}{\sigma_g^2}\boldsymbol{\beta}^T\boldsymbol{\beta}\end{aligned}$$

which is the penalised regression form for ridge regression with $\lambda = \frac{\sigma^2}{\sigma_g^2}$. The similarity between random normal models and ridge regression was first shown by Lindley & Smith (1972).

The LASSO model can likewise be defined. In particular the model effects are assumed to be independently distributed as double exponential variables with mean zero and variance $2\phi^2$. The functional form for the centred double exponential distribution is

$$f(\beta_i) = \frac{1}{2\phi}\exp\left(-\frac{1}{\phi}|\beta_i|\right).$$

The moment generating function for the double exponential distribution is given in Theorem A.7 and its variance is given in Corollary A.1. The functional form of the log-joint distribution, ignoring terms not involving $\boldsymbol{\beta}$, is

$$\log f(\boldsymbol{y}, \boldsymbol{\beta}) = (\boldsymbol{y} - \boldsymbol{L\beta})^T(\boldsymbol{y} - \boldsymbol{L\beta}) + \frac{\sigma^2}{\phi}\sum_{i=1}^{q}|\beta_i|$$

which is the penalised regression for the LASSO with $\lambda = \sigma^2/\phi$.

These results imply that the estimates from a penalised likelihood, or equivalently a constrained likelihood, can be identical to those from a random effects model. However, the random effects model specification provides insight into the meaning of the penalty parameter. It is the ratio of the residual dispersion parameter and the dispersion parameter of the random effects. If these are known *a priori* then the choice of penalty parameter and constraint is immediate. If the dispersion parameters are unknown then they could be estimated. This provides a method for choosing the penalty and implies that the penalty has an interpretation.

Other penalised regressions can be expressed as a random effects model, for example a subset of the bridge models (Fu, 1998) and the elastic net (Zou & Hastie, 2005). These cases may require the random effects to have a distribution that is not common, involves variance and shape parameters or is a mixture of distributions.

4.6 Computation for Ridge Regression

If the model postulated contains only terms treated as ridge regression terms, then the estimate of $\boldsymbol{\beta}$ can be directly found by optimising the penalised regression. The same result also follows from treating the ridge effects as random normal variates. With $\lambda = \frac{\sigma^2}{\sigma_g^2}$ known, both approaches give

$$\tilde{\boldsymbol{\beta}} = (\boldsymbol{L}^T\boldsymbol{L} + \lambda \boldsymbol{I}_q)^{-1}\boldsymbol{L}\boldsymbol{y}. \tag{4.6.1}$$

If there are fixed (unconstrained) effects then the variation in the observed outcomes caused by them has to be accounted for. The most efficient method to do this is joint estimation via the mixed model equations in Chapter 3. Assuming that there is no residual variance structure and that the specification of the normal variates (ridge parameters) is independent and identically distributed, then the reduced MME can be given

$$\begin{bmatrix} \boldsymbol{X}^T\boldsymbol{X} & \boldsymbol{X}^T\boldsymbol{L} \\ \boldsymbol{L}^T\boldsymbol{X} & \boldsymbol{L}^T\boldsymbol{L} + \lambda\boldsymbol{I} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\tau}} \\ \tilde{\boldsymbol{\beta}} \end{bmatrix} = \begin{bmatrix} \boldsymbol{X}^T\boldsymbol{y} \\ \boldsymbol{L}^T\boldsymbol{y} \end{bmatrix}.$$

It is often suggested that the outcomes should be pre-adjusted for fixed effects by subtracting from the outcomes predictions based on a fixed effect only model. This is not advised as pre-adjusting imposes an unwanted covariance structure on the adjusted outcomes. For illustration, assume a model with fixed effects $\boldsymbol{\tau}$ and associated design matrix $\boldsymbol{X}$. The adjusted outcomes (or residuals) from the fixed effect only model are

$$\begin{aligned} \tilde{\boldsymbol{e}} &= \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\tau}} \\ &= \left(\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\right)\boldsymbol{y}. \end{aligned}$$

The mean and variance of $\tilde{\boldsymbol{e}}$ are

$$\begin{aligned} \mathrm{E}\left(\tilde{\boldsymbol{e}}\right) &= \left(\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\right)\mathrm{E}\left(\boldsymbol{y}\right) \\ &= \boldsymbol{0} \\ \mathrm{var}\left(\tilde{\boldsymbol{e}}\right) &= \left(\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\right)\mathrm{var}\left(\boldsymbol{y}\right)\left(\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\right)^T \\ &= \sigma^2\left(\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\right). \end{aligned}$$

Those who propose pre-adjusting often assume that $\text{var}\,(\tilde{\boldsymbol{e}})$ is homogeneous and diagonal but this assumption is clearly not correct for the general case.

A special case is when there is only an overall mean parameter in the fixed effects. Then

$$\text{var}\,(\tilde{\boldsymbol{e}}) = \sigma^2(\boldsymbol{I} - \frac{1}{n}\boldsymbol{J})$$

where $\boldsymbol{J}$ is a $n \times n$ matrix of ones. If n is large then preadjusting the outcomes will be approximately correct. However, it should still be avoided if possible.

In most practical situations the penalty parameter, equivalently the constraint parameter or the variance components, are unknown prior to analysis. This means that they have to be estimated. A common method is to use cross-validation or some variant of it to choose the penalty parameter that minimises the prediction error. An alternative approach is to estimate the variance components, preferably via REML, and to use these estimates to solve the optimisation problem. The two approaches will not, in general, give identical results.

4.7 Computation for the LASSO

All the methods of estimation known to the author require the explanatory variables to be standardised, as constrained regression is not invariant to arbitrary changes of scale. Also, the outcomes have to be standardised or at least centred for known fixed effects before estimation of the LASSO effects. As mentioned in Section 4.6 this is potentially misleading as the adjusted outcomes are not independent. The method for computation presented here is the algorithm of Osborne et al. (2000) and is based on solving the constraint problem. Other methods to solve the constraint problem include the quadratic programming solution in Tibshirani (1996), the iterative ridge approximation method in Öjelund et al. (2001), the linear constraint method of Miller (2002) and the least angle regression method in Efron et al. (2004). The least angle regression method is a stylised form of the algorithm of Osborne et al. (2000). There is also a method to estimate LASSO effects from the penalised regression (Fu, 1998). These should all give the same answers for equivalent constraint and penalty parameters.

One of the attractive attributes of the algorithm of Osborne et al. (2000) is that it works even if the number of explanatory variables is greater than the number of observations. Only the LAR algorithm of Efron et al. (2004) shares this characteristic. The other algorithms either start at the full least squares estimate or use them in the calculations.

4.7.1 *First Order Conditions*

Commonly, optimisation is carried out by taking derivatives and finding the critical points of the objective function. This is not possible for the LASSO as the Lagrangian of the constraint problem (4.5.1) is not differentiable. Instead a generalisation of the differential, called the sub-differential, is used (e.g. Osborne, 1985). It is defined on convex functions and is stated formally in Definition C.3. The sub-differential of the Lagrangian for the LASSO problem is

$$\partial_{\boldsymbol{\beta}}\mathscr{L}(\boldsymbol{\beta},\lambda) = -\boldsymbol{L}^T(\boldsymbol{y}-\boldsymbol{L}\boldsymbol{\beta}) + \lambda\boldsymbol{v} \tag{4.7.1}$$

where $\boldsymbol{v} = (v_1, v_2, \ldots, v_q)^T$ is chosen such that $v_i = -1$ if $\beta_i < 0$, $v_i = +1$ is $\beta_i > 0$ and $v_i \in (-1, 1)$ if $\beta_i = 0$. This form implies that $\boldsymbol{v}^T\boldsymbol{\beta} = \|\boldsymbol{\beta}\|_1$ and $\|\bar{\boldsymbol{v}}\|_\infty = 1$. The vector $\boldsymbol{v}$ is introduced to allow for the different possibilities of sub-gradient (see Definition C.2) when elements of the random variable $\boldsymbol{\beta}$ are zero. In a linear programming analogy the $\boldsymbol{v}$ behaves in a manner similar to slack variables.

At an optimal point (a solution), the zero vector must be a member of the sub-differential (e.g. Osborne, 1985). Let $\bar{\boldsymbol{\beta}}$ be the optimal point and let $\bar{\boldsymbol{v}}$ correspond to this solution. Then

$$\mathbf{0} = -\boldsymbol{L}^T(\boldsymbol{y}-\boldsymbol{L}\bar{\boldsymbol{\beta}}) + \lambda\bar{\boldsymbol{v}} \tag{4.7.2}$$

and hence

$$\bar{\boldsymbol{\beta}} = (\boldsymbol{L}^T\boldsymbol{L})^{-}\boldsymbol{L}^T\boldsymbol{y} - \lambda(\boldsymbol{L}^T\boldsymbol{L})^{-}\bar{\boldsymbol{v}}.$$

where $\boldsymbol{D}^-$ is a generalised inverse of $\boldsymbol{D}$.

If the L_∞ norm of (4.7.2) is taken then $\lambda = \|\boldsymbol{L}^T(\boldsymbol{y}-\boldsymbol{L}\bar{\boldsymbol{\beta}})\|_\infty$ and so the optimal solution $\bar{\boldsymbol{v}}$ is given by

$$\bar{\boldsymbol{v}} = \frac{\boldsymbol{L}^T(\boldsymbol{y}-\boldsymbol{L}\bar{\boldsymbol{\beta}})}{\|\boldsymbol{L}^T(\boldsymbol{y}-\boldsymbol{L}\bar{\boldsymbol{\beta}})\|_\infty}. \tag{4.7.3}$$

This expression gives an optimality check. If $\bar{\boldsymbol{\beta}}$ is not optimal then $\bar{\boldsymbol{v}}$ will have elements outside of $[-1, 1]$. This arises as $\mathbf{0} \notin \partial_{\boldsymbol{\beta}}\mathscr{L}(\boldsymbol{\beta},\lambda)$. Hence, the maximiser $\bar{\boldsymbol{\beta}}$ is the value which has a corresponding $\bar{\boldsymbol{v}}$ with elements within $[-1, 1]$ (Osborne et al., 2000).

4.7.2 *Linearisation and Optimal Descent Directions*

The sub-differential of the Lagrangian for the LASSO problem (4.7.1) is differentiable with respect to $\boldsymbol{\beta}$ and λ, abusing notation call this quantity $\nabla^2\mathscr{L}(\boldsymbol{\beta},\lambda)$. This implies that methods requiring the sub-differential's derivative to solve the sub-differential can be used. An obvious choice is the iterative Newton-Raphson method which updates the solution by taking linear steps from the current solution. Let the updated solution be related to the previous solution by

$$\begin{pmatrix} \boldsymbol{\beta}^{(k+1)} \\ \lambda^{(k+1)} \end{pmatrix} = \begin{pmatrix} \boldsymbol{\beta}^{(k)} \\ \lambda^{(k)} \end{pmatrix} + \begin{pmatrix} \boldsymbol{h}^{(k)} \\ \nu^{(k)} \end{pmatrix}.$$

The first order Taylor series around the previous solutions, so that the updated solutions solve the approximation, give the equations for choosing $\boldsymbol{h}^{(k)}$ and $\nu^{(k)}$ (Osborne et al., 2000).

They are

$$\nabla^2 \mathscr{L}(\boldsymbol{\beta}^{(k)}, \lambda^{(k)}) \begin{pmatrix} \boldsymbol{h}^{(k)} \\ \nu^{(k)} \end{pmatrix} = \partial_{(\boldsymbol{\beta},\lambda)} \mathscr{L}(\boldsymbol{\beta}^{(k)}, \lambda^{(k)}).$$

The equations for choosing the linear descent directions for LASSO estimation based on these general ones are

$$\begin{pmatrix} \boldsymbol{L}^T \boldsymbol{L} & \bar{\boldsymbol{v}}^{(k)} \\ \bar{\boldsymbol{v}}^{(k)^T} & \boldsymbol{0} \end{pmatrix} \begin{pmatrix} \boldsymbol{h}^{(k)} \\ \nu^{(k)} \end{pmatrix} = \begin{pmatrix} \boldsymbol{L}^T(\boldsymbol{y} - \boldsymbol{L}\bar{\boldsymbol{\beta}}^{(k)}) \\ t - \bar{\boldsymbol{v}}^{(k)^T} \bar{\boldsymbol{\beta}}^{(k)} \end{pmatrix}. \tag{4.7.4}$$

These equations specify the linear form of the Karush-Kuhn-Tucker conditions in Result C.1. The choice of descent direction chosen by these equations will be the optimal choice from the current position.

4.7.3 *Interior Point Descent Algorithm*

The interior point descent algorithm of Osborne et al. (2000) is based on this Newton-Raphson iterative process described for a single updating step in (4.7.4). The algorithm starts with an arbitrary interior point, uses the steps defined by $\boldsymbol{h}^{(k)}$ to find the next point and checks to see if this point is the optimal solution via (4.7.3). However, this algorithm chooses descent directions only for a subset of $\{\beta_i\}$ - those which are non-zero at the current iteration. This is done to ensure convergence, even when the number of explanatory variables is larger than the number of observations.

Let $\mathcal{S} = \{i_1, i_2, \ldots, i_{q_\mathcal{S}}\}$ be the set of $q_\mathcal{S}$ indices for which $\bar{v}_{i_j} = \pm 1$, where $i_j \in \mathcal{S}$. The set $\mathcal{S}$ is called the active set and corresponds to those explanatory variables whose effect estimates are non-zero. Without loss of generality, define $\boldsymbol{\beta} = (\boldsymbol{\beta}_\mathcal{S}^T, \boldsymbol{0}^T)^T$, $\boldsymbol{L} = (\boldsymbol{L}_\mathcal{S}, \boldsymbol{L}_\mathcal{Z})$ and $\boldsymbol{h} = (\boldsymbol{h}_\mathcal{S}^T, \boldsymbol{0}^T)^T$, that is a rearrangement of these entities so that the elements with index in $\mathcal{S}$ occur first. Also let $\boldsymbol{\theta}^T = (\boldsymbol{\theta}_\mathcal{S}^T, \boldsymbol{0}^T) = (\text{sign}(\bar{\boldsymbol{\beta}}_\mathcal{S}^{(k)})^T, \boldsymbol{0}^T)$ be the vector of current signs of the solutions augmented with zeros. The descent algorithm performs a local linearisation of the LASSO constraint problem around the current $\bar{\boldsymbol{\beta}}_\mathcal{S}^{(k)}$

$$\begin{aligned} \underset{\boldsymbol{h}^{(k)}}{\text{minimise:}} \quad & \left(\boldsymbol{y} - \boldsymbol{L}_\mathcal{S}(\bar{\boldsymbol{\beta}}_\mathcal{S}^{(k)} + \boldsymbol{h}_\mathcal{S}^{(k)})\right)^T \left(\boldsymbol{y} - \boldsymbol{L}_\mathcal{S}(\bar{\boldsymbol{\beta}}_\mathcal{S}^{(k)} + \boldsymbol{h}_\mathcal{S}^{(k)})\right) \\ \text{subject to:} \quad & \theta_\mathcal{S}^{(k)^T} \left(\bar{\boldsymbol{\beta}}_\mathcal{S}^{(k)} + \boldsymbol{h}_\mathcal{S}^{(k)}\right) \leq t. \end{aligned}$$

If the constraint is active then the choice of $\boldsymbol{h}_\mathcal{S}$ can be given by the linear Karush-Kuhn-Tucker conditions in (4.7.4) reduced to the set $\mathcal{S}$. That is

$$\begin{pmatrix} \boldsymbol{L}_\mathcal{S}^T \boldsymbol{L}_\mathcal{S} & \boldsymbol{\theta}_\mathcal{S} \\ \boldsymbol{\theta}_\mathcal{S}^T & \boldsymbol{0} \end{pmatrix} \begin{pmatrix} \boldsymbol{h}_\mathcal{S}^{(k)} \\ \nu^{(k)} \end{pmatrix} = \begin{pmatrix} \boldsymbol{L}_\mathcal{S}^T(\boldsymbol{y} - \boldsymbol{L}_\mathcal{S}\bar{\boldsymbol{\beta}}_\mathcal{S}^{(k)}) \\ t - \boldsymbol{\theta}_\mathcal{S}^T \bar{\boldsymbol{\beta}}_\mathcal{S}^{(k)} \end{pmatrix}. \tag{4.7.5}$$

Solutions to these equations are

$$\begin{aligned} \boldsymbol{h}_\mathcal{S}^{(k)} &= (\boldsymbol{L}_\mathcal{S}^T \boldsymbol{L}_\mathcal{S})^{-1}(\boldsymbol{L}_\mathcal{S}^T(\boldsymbol{y} - \boldsymbol{L}_\mathcal{S}\bar{\boldsymbol{\beta}}_\mathcal{S}^{(k)}) - \nu^{(k)}\boldsymbol{\theta}_\mathcal{S}) \qquad \text{where} \\ \nu^{(k)} &= \max\left(0, \frac{\boldsymbol{\theta}_\mathcal{S}^T(\boldsymbol{L}_\mathcal{S}^T \boldsymbol{L})^{-1}\boldsymbol{L}_\mathcal{S}^T \boldsymbol{y} - t}{\boldsymbol{\theta}_\mathcal{S}^T(\boldsymbol{L}_\mathcal{S}^T \boldsymbol{L}_\mathcal{S})^{-1}\boldsymbol{\theta}_\mathcal{S}}\right). \end{aligned}$$

The increment of $\nu^{(k)}$ is forced to be non-negative. If it were not then the increment for the estimates could produce new estimates equal to or greater than the least squares estimates, the unconstrained problem on a subset of explanatory variables.

This choice of $\boldsymbol{h}_{\mathcal{S}}^{(k)}$ allows the sign of any element of $\bar{\boldsymbol{\beta}}_{\mathcal{S}}^{(k+1)}$ to change. If this occurs then convergence is not guaranteed. Osborne et al. (2000) force the estimate updates to be of the same sign as the previous estimates. If this is the case then the updates are said to be sign feasible. If the updates are not sign feasible then the updating step is reduced so that the offending element is identically zero; it is removed from the active set $\mathcal{S}$.

After each updating step the current solution is checked for optimality via (4.7.3). If it is optimal the algorithm stops, if not then another variable is added to the active set. This variable is chosen as the one with maximum corresponding v_i. The algorithm is now ready to be stated in full and is based on the presentation in Lokhorst (1999). It is also given in the flow diagram in Figure 4.2.

1. **Main Step** To move from iteration k to iteration $(k+1)$ solve the local optimisation problem specified in (4.7.5) to obtain the solutions for $\boldsymbol{h}_{\mathcal{S}}^{(k)}$ and $\nu^{(k)}$. Consider the updated parameter solutions $\bar{\boldsymbol{\beta}}_{\mathcal{S}}^{(k+1)} = \bar{\boldsymbol{\beta}}_{\mathcal{S}}^{(k)} + \boldsymbol{h}_{\mathcal{S}}^{(k)}$.

2. **Sub-routine**

   - **IF** $\bar{\boldsymbol{\beta}}^{(k+1)}$ is not sign feasible then

     (a) Find the smallest $\gamma \in (0,1)$ for which there exists a $i \in \mathcal{S}$ such that $\bar{\beta}_i^{(k)} + \gamma h_i^{(k)} = 0$. This moves the *first new zero component* to zero in the descent direction.

     (b) Update the active set $\mathcal{S}$ by removing i and recompute solutions to (4.7.5) with the reduced system.

     (c) Iterate until the new solution is sign feasible.

   - **ELSE** $\bar{\boldsymbol{\beta}}^{(k+1)}$ is sign feasible.

     (a) Calculate $\bar{\boldsymbol{v}}^{(k+1)}$ as per (4.7.3).

       – If $|\bar{v}_j^{(k+1)}| = |\theta_j^{(k+1)}|$ for $j \in \mathcal{S}$ and $-1 < \bar{v}_j^{(k+1)} < 1$ for $j \notin \mathcal{S}$ then $\bar{\boldsymbol{\beta}}^{(k+1)}$ is a solution and algorithm is stopped.

       – Else find $j \notin \mathcal{S}$ such that $\bar{v}_j^{(k+1)}$ has maximum absolute value. Update $\mathcal{S}$, $\bar{\boldsymbol{\beta}}_{\mathcal{S}}^{(k+1)}$ and $\boldsymbol{\theta}_{\mathcal{S}}^{(k+1)}$ by appending j, 0 and $\text{sign}(\bar{v}^{(k+1)})$ respectively. Return to the main step and start process again.

Convergence is guaranteed as each updated solution is within the solution space and arises from a descent step which reduces the objective function (Osborne et al., 2000). From a geometrical viewpoint, each update moves the current solution (of dimension $|\mathcal{S}|$) to a point within the solution space which decreases the objective function. This update decreases the objective function the most, given the updated solution is of dimension $|\mathcal{S}|$. The dimension is then increased by one and the whole process continues. Eventually, after a finite number

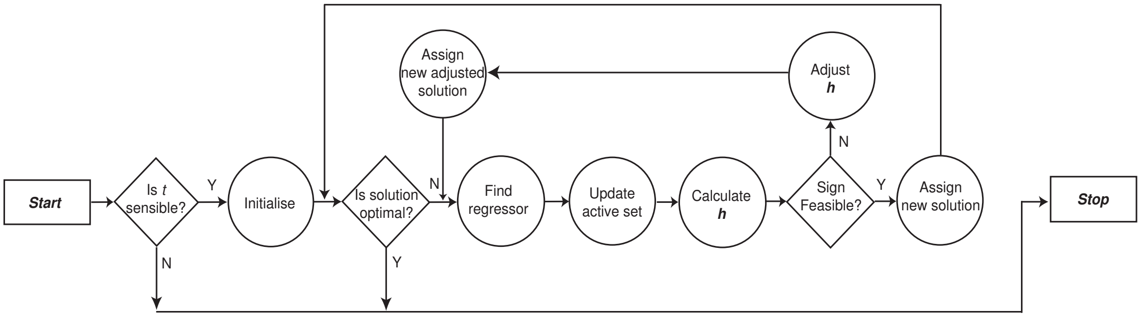

Figure 4.2: Flow diagram for interior point descent algorithm.

of steps, the current solution is on the boundary of the solution space - the L_1-sphere. The update then moves the current solution around the boundary until the optimal solution is found. This is efficient as linear steps move the solution along the edges of the polyhedral solution space. Convergence is guaranteed in a finite number of iterations as there are only finitely many combinations of the elements of $\boldsymbol{\beta}$.

For any particular data set, the need may arise for estimating the model for multiple values of t. The efficiency of estimation in this set of models can be increased if the constraints are ordered so that the model with the smallest constraint is estimated first and then the next smallest and so on. The estimates from the model with the smaller constraint can then be used as the starting point for the model with the larger constraint. Further, many of the quantities needed, such as $(\boldsymbol{L}_{\mathcal{S}}^T\boldsymbol{L}_{\mathcal{S}})^{-1}$, can be used in the first iteration of the bigger model.

Starting the iterative procedure with $\mathcal{S} = \emptyset$ shows some similarities with more standard model building methods (Osborne et al., 2000). In particular, if the initial active set is $\mathcal{S} = \emptyset$ then each iteration corresponds to the addition (or deletion) of a new variable, in this sense it is a forward moving algorithm (Section 4.3). Unlike the forward stepwise algorithm, the interior point algorithm is governed by a global optimality criterion.

4.8 Biased Estimate Standard Errors

4.8.1 *Ridge Regression*

The variance of the ridge regression estimates can be deduced from its function form. If the estimates (4.6.1) are a linear combination of the normal outcome data then they will also have a normal distribution with mean $\mathbf{0}$ and variance

$$\begin{aligned}\operatorname{var}\left(\tilde{\boldsymbol{\beta}}\right) &= \operatorname{var}\left((\boldsymbol{L}^T\boldsymbol{L} + \lambda\boldsymbol{I}_q)^{-1}\boldsymbol{L}\boldsymbol{y}\right) \\ &= (\boldsymbol{L}^T\boldsymbol{L} + \lambda\boldsymbol{I}_q)^{-1}\boldsymbol{L}\operatorname{var}(\boldsymbol{y})\,\boldsymbol{L}^T(\boldsymbol{L}^T\boldsymbol{L} + \lambda\boldsymbol{I}_q)^{-1}. \qquad (4.8.1)\end{aligned}$$

Commonly $\operatorname{var}(\boldsymbol{y})$ is assumed to be diagonal and homogeneous. However, the random effects model formulation suggests that this is not appropriate.

4.8.2 *The LASSO*

Approximate standard errors for the LASSO have been proposed in Tibshirani (1996) and Osborne et al. (2000). Both are based on the same idea; approximating the LASSO estimates by ridge regression estimates and then using (4.8.1). Both approximations correspond to using an approximation for the L_1-norm (Osborne et al., 2000).

The approximation proposed by Tibshirani (1996) reformulates the constraint as

$$\sum_{i=1}^{q} |\beta_i| = \sum_{i=1}^{q} \frac{\beta_j^2}{|\beta_j|}.$$

This enables an approximation of the form of the ridge regression estimates of $\bar{\boldsymbol{\beta}} = (\boldsymbol{L}^T\boldsymbol{L} + \lambda\boldsymbol{W}^-)^{-1}\boldsymbol{L}^T\boldsymbol{y}$, where $\boldsymbol{W}$ is a diagonal matrix with the j^th diagonal element being $|\bar{\beta}_j|$, $\boldsymbol{W}^-$ is a generalised inverse of $\boldsymbol{W}$ and λ is the Lagrange multiplier. Based on (4.8.1) and assuming that $\text{var}\,(\boldsymbol{y}) = \sigma^2\boldsymbol{I}$ the approximate variance is

$$\text{var}\left(\bar{\boldsymbol{\beta}}\right) \approx \sigma^2(\boldsymbol{L}^T\boldsymbol{L} + \lambda\boldsymbol{W}^-)^{-1}\boldsymbol{L}\boldsymbol{L}^T(\boldsymbol{L}^T\boldsymbol{L} + \lambda\boldsymbol{W}^-)^{-1}.$$

This approximation gives an estimated variance of 0 for estimates with $\bar{\beta}_j = 0$ (Tibshirani, 1996 and Osborne et al., 2000). This is cause for concern. The numerical values obtained from this equation are also dependent on the choice of generalised inverse.

To overcome this problem Osborne et al. (2000) suggested a different approximation. They note that at a solution (4.7.2) the Lagrange multiplier must have the form

$$\lambda = \frac{(\boldsymbol{y} - \boldsymbol{L}\bar{\boldsymbol{\beta}})^T\boldsymbol{L}\bar{\boldsymbol{\beta}}}{\|\bar{\boldsymbol{\beta}}\|_1}$$

and the form of $\bar{\boldsymbol{v}}$ is given in (4.7.3). At the solution $\boldsymbol{0} \in \partial_{\boldsymbol{\beta}}\mathscr{L}(\boldsymbol{\beta}\lambda)$ and so from (4.7.2)

$$\begin{aligned}
\boldsymbol{L}^T\boldsymbol{y} &= \boldsymbol{L}^T\boldsymbol{L}\bar{\boldsymbol{\beta}} + \lambda\bar{\boldsymbol{v}} \\
&= \left(\boldsymbol{L}^T\boldsymbol{L} + \frac{\boldsymbol{L}^T(\boldsymbol{y} - \boldsymbol{L}\bar{\boldsymbol{\beta}})(\boldsymbol{y} - \boldsymbol{L}\bar{\boldsymbol{\beta}})^T\boldsymbol{L}}{\|\boldsymbol{L}^T(\boldsymbol{y} - \boldsymbol{L}\bar{\boldsymbol{\beta}})\|_\infty\|\bar{\boldsymbol{\beta}}\|_1}\right)\bar{\boldsymbol{\beta}} \\
&= (\boldsymbol{L}^T\boldsymbol{L} + \boldsymbol{W}_o)\bar{\boldsymbol{\beta}} \qquad\qquad \text{say.}
\end{aligned}$$

This gives the LASSO estimates to be

$$\bar{\boldsymbol{\beta}} = (\boldsymbol{L}^T\boldsymbol{L} + \boldsymbol{W}_o)^{-1}\boldsymbol{L}^T\boldsymbol{y}.$$

The approximate variance of this estimator is

$$\begin{aligned}
\text{var}\left(\bar{\boldsymbol{\beta}}\right) &\approx (\boldsymbol{L}^T\boldsymbol{L} + \boldsymbol{W}_o)^{-1}\boldsymbol{L}^T\text{var}\,(\boldsymbol{y})\,\boldsymbol{L}(\boldsymbol{L}^T\boldsymbol{L} + \boldsymbol{W}_o)^{-1} \\
&= \sigma^2(\boldsymbol{L}^T\boldsymbol{L} + \boldsymbol{W}_o)^{-1}\boldsymbol{L}^T\boldsymbol{L}(\boldsymbol{L}^T\boldsymbol{L} + \boldsymbol{W}_o)^{-1}
\end{aligned}$$

if $\text{var}\,(\boldsymbol{y}) = \sigma^2\boldsymbol{I}$.

The usefulness of these variance estimators is questioned as the distribution of the estimates will typically have a mass of probability at zero (Osborne et al., 2000). This observation is formally investigated from an asymptotic viewpoint in Knight & Fu (2000).

Chapter 5

Some Results for the LASSO Random Model

5.1 Introduction

Under the random effects model formulation for the LASSO, the LASSO estimates are defined to be the mode of the joint distribution of the observed outcomes $\boldsymbol{y}$ and the random effects $\boldsymbol{\beta}$ (Tibshirani, 1996). They are modes, not means, of the joint distribution $f(\boldsymbol{y}, \boldsymbol{\beta})$ and hence also of the predictive distribution (or posterior distribution in Bayesian terminology) $f(\boldsymbol{\beta}|\boldsymbol{y}) = f(\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}})$. In this chapter some of the important distributions related to the LASSO random effects model are investigated along with some of the moments of these distributions. This is performed initially using a simplistic model with only one LASSO random effect. The multiple LASSO effect case is intractable. However, the marginal distribution is available.

The LASSO estimate is compared to the mean of the predictive distribution. They are similar in many situations, except when the mean and mode approach zero. In such situations the mean, like the ridge regression estimate, will never be identically zero. This implies that the *selection* attribute of the LASSO rests on the estimates being modes.

Lastly, in this chapter, it is shown that the non-zero LASSO estimates $\bar{\boldsymbol{\beta}}_{\mathcal{S}}$ can be estimated free of the zero estimates $\bar{\boldsymbol{\beta}}_{\mathcal{Z}}$. This suggests that computation can be reduced significantly *if* the LASSO effects are partitioned into those whose estimate is non-zero and those whose estimate is zero prior to calculating the effect.

5.2 Random Effects Model with Single Effect

Initially consider the simplest (pathological) random effects model, with only one effect β

$$\boldsymbol{y} = \boldsymbol{l}\beta + \boldsymbol{e} \tag{5.2.1}$$

where $\boldsymbol{y}$ is a $n \times 1$ vector of observed random variables, $\boldsymbol{l}$ is the design vector for the random effect $\beta \sim \mathrm{DE}(0, 2\phi^2)$ and $\boldsymbol{e}$ is the vector of independent residuals whose distribution is assumed normal with mean zero and common variance σ^2.

5.2.1 *Marginal Distribution of Observations*

The marginal distribution is important in mixed model theory as it forms the basis for the likelihood. In the single random effect LASSO model (5.2.1) the marginal distribution is found by partitioning the interval of integration into positive and negative real numbers. This circumvents the problem of integrating the absolute value function. The resulting integral is a function of the LASSO estimate assuming that the estimate were positive or negative. If the LASSO estimate has equal probability of being either positive or negative then the marginal likelihood will be dominated by the corresponding term.

Theorem 5.1. *The marginal distribution for the random effects model in* (5.2.1) *is*

$$f(\boldsymbol{y}) = \frac{(2\pi\sigma^2_{\hat{\beta}})^{\frac{1}{2}}}{2\phi(2\pi\sigma^2)^{\frac{n}{2}}}\exp\left(-\frac{\boldsymbol{y}^T\boldsymbol{y}}{2\sigma^2}\right) \times \\ \left[\exp\left(\frac{\bar{\beta}^2_+}{2\sigma^2_{\hat{\beta}}}\right)\left(1-\Phi\left(-\frac{\bar{\beta}_+}{\sigma_{\hat{\beta}}}\right)\right) + \exp\left(\frac{\bar{\beta}^2_-}{2\sigma^2_{\hat{\beta}}}\right)\Phi\left(-\frac{\bar{\beta}_-}{\sigma_{\hat{\beta}}}\right)\right] \tag{5.2.2}$$

where $\hat{\beta} = (\boldsymbol{l}^T\boldsymbol{l})^{-1}\boldsymbol{l}^T\boldsymbol{y}$, $\sigma^2_{\hat{\beta}} = \sigma^2(\boldsymbol{l}^T\boldsymbol{l})^{-1}$, $\bar{\beta}_+ = \hat{\beta} - \sigma^2_{\hat{\beta}}/\phi$ *and* $\bar{\beta}_- = \hat{\beta} + \sigma^2_{\hat{\beta}}/\phi$. *Note that* $\bar{\beta}_+$ *is the LASSO estimate if* β *is known to be positive, likewise* $\bar{\beta}_-$ *is the LASSO estimate if* β *is known to be negative.*

Proof. The marginal distribution is found by integrating over the random effect β

$$\begin{aligned}
f(\boldsymbol{y}) &= \int_{-\infty}^{\infty} f(\boldsymbol{y}|\beta)f(\beta)d\beta \\
&= \frac{1}{2\phi(2\pi\sigma^2)^{\frac{n}{2}}}\int_{-\infty}^{\infty}\exp\left(-\frac{1}{2\sigma^2}(\boldsymbol{y}-\boldsymbol{l}\beta)^T(\boldsymbol{y}-\boldsymbol{l}\beta) - \frac{1}{\phi}|\beta|\right)d\beta \\
&= \frac{1}{2\phi(2\pi\sigma^2)^{\frac{n}{2}}}\int_{-\infty}^{\infty}\exp\left(-\frac{\boldsymbol{y}^T\boldsymbol{y}}{2\sigma^2} - \frac{\boldsymbol{l}^T\boldsymbol{l}}{2\sigma^2}\left(\beta^2 - 2(\boldsymbol{l}^T\boldsymbol{l})^{-1}\boldsymbol{l}^T\boldsymbol{y}\beta + \frac{2\sigma^2(\boldsymbol{l}^T\boldsymbol{l})^{-1}}{\phi}|\beta|\right)\right)d\beta \\
&= \frac{1}{2\phi(2\pi\sigma^2)^{\frac{n}{2}}}\exp\left(-\frac{\boldsymbol{y}^T\boldsymbol{y}}{2\sigma^2}\right)\left[\int_0^{\infty}\exp\left(-\frac{1}{2\sigma^2_{\hat{\beta}}}\left(\beta^2 - 2\bar{\beta}_+ + \bar{\beta}^2_+\right) + \frac{\bar{\beta}^2_+}{2\sigma^2_{\hat{\beta}}}\right)d\beta\right. \\
&\qquad \left. + \int_{-\infty}^{0}\exp\left(-\frac{1}{2\sigma^2_{\hat{\beta}}}\left(\beta^2 - 2\bar{\beta}_- + \bar{\beta}^2_-\right) + \frac{\bar{\beta}^2_-}{2\sigma^2_{\hat{\beta}}}\right)d\beta\right] \\
&= \frac{1}{2\phi(2\pi\sigma^2)^{\frac{n}{2}}}\exp\left(-\frac{\boldsymbol{y}^T\boldsymbol{y}}{2\sigma^2}\right)\left[\exp\left(\frac{\bar{\beta}^2_+}{2\sigma^2_{\hat{\beta}}}\right)\int_0^{\infty}\exp\left(-\frac{1}{2\sigma^2_{\hat{\beta}}}\left(\beta-\bar{\beta}_+\right)^2\right)d\beta\right. \\
&\qquad \left. +\exp\left(\frac{\bar{\beta}^2_-}{2\sigma^2_{\hat{\beta}}}\right)\int_{-\infty}^{0}\exp\left(-\frac{1}{2\sigma^2_{\hat{\beta}}}\left(\beta-\bar{\beta}_-\right)^2\right)d\beta\right] \\
&= \frac{1}{2\phi(2\pi\sigma^2)^{\frac{n}{2}}}\exp\left(-\frac{\boldsymbol{y}^T\boldsymbol{y}}{2\sigma^2}\right)\left[\exp\left(\frac{\bar{\beta}^2_+}{2\sigma^2_{\hat{\beta}}}\right)A + \exp\left(\frac{\bar{\beta}^2_-}{2\sigma^2_{\hat{\beta}}}\right)B\right] \qquad \text{say.}
\end{aligned} \tag{5.2.3}$$

Initially consider the integral A. Change the variable (e.g. Apostol, 1967) using

$$u = \frac{\beta - \bar{\beta}_+}{\sigma_{\hat{\beta}}} \qquad \text{giving} \qquad \frac{du}{d\beta} = \frac{1}{\sigma_{\hat{\beta}}}.$$

This gives

$$\begin{aligned} A &= \sqrt{2\pi\sigma^2_{\hat{\beta}}} \int_{-\frac{-\bar{\beta}_+}{\sigma_{\hat{\beta}}}}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}u^2\right) du \\ &= \sqrt{2\pi\sigma^2_{\hat{\beta}}} \left(1 - \Phi\left(-\frac{\bar{\beta}_+}{\sigma_{\hat{\beta}}}\right)\right). \end{aligned}$$

Similarly

$$B = \sqrt{2\pi\sigma^2_{\hat{\beta}}} \Phi\left(-\frac{\bar{\beta}_-}{\sigma_{\hat{\beta}}}\right).$$

The result is obtained by substituting A and B into (5.2.3). □

When the LASSO estimate is large with respect to (w.r.t.) $\sigma_{\hat{\beta}}$ and positive then the first term of the marginal likelihood will be dominant and the second almost zero. Exactly the opposite occurs when the LASSO estimate is large w.r.t. $\sigma_{\hat{\beta}}$ and negative. If neither $\bar{\beta}_+/\sigma_{\hat{\beta}}$ nor $\bar{\beta}_-/\sigma_{\hat{\beta}}$ is large then the marginal distribution will not be dominated by either term.

It will be useful for the derivation of the expected value of the posterior distribution to have an alternative form for the marginal distribution.

Corollary 5.1. *The marginal distribution for the random effects model* (5.2.1) *can be expressed in terms of the least squares estimate.*

$$\begin{aligned} f(\boldsymbol{y}) = \frac{(2\pi\sigma^2_{\hat{\beta}})^{\frac{1}{2}}}{2\phi(2\pi\sigma^2)^{\frac{n}{2}}} \exp\left(-\frac{\boldsymbol{y}^T\boldsymbol{y}}{2\sigma^2}\right) \exp\left(\frac{\hat{\beta}^2}{2\sigma^2_{\hat{\beta}}} + \frac{\sigma^2_{\hat{\beta}}}{2\phi^2}\right) \times \\ \left[\exp\left(-\frac{\hat{\beta}}{\phi}\right)\left(1 - \Phi\left(-\frac{\bar{\beta}_+}{\sigma_{\hat{\beta}}}\right)\right) + \exp\left(\frac{\hat{\beta}}{\phi}\right)\Phi\left(-\frac{\bar{\beta}_-}{\sigma_{\hat{\beta}}}\right)\right] \end{aligned} \qquad (5.2.4)$$

Proof. Substitute $\bar{\beta}_+ = \hat{\beta} - \frac{\sigma^2_{\hat{\beta}}}{\phi}$ and $\bar{\beta}_- = \hat{\beta} + \frac{\sigma^2_{\hat{\beta}}}{\phi}$ and the result is obtained immediately. □

5.2.2 *Predictive Distribution*

Consider that the dispersion parameters are known or that there are plug-in estimates available. The predicted effects for the random effects model are often taken as a summary of the predictive distribution, often this summary is the mean or mode of the distribution.

Theorem 5.2. *The predictive distribution of the scalar effect β is*

$$f(\beta|\boldsymbol{y}) = \frac{\exp\left(\frac{1}{\sigma^2}\boldsymbol{y}^T\boldsymbol{l}\beta - \frac{1}{2\sigma^2_{\hat{\beta}}}\beta^2 - \frac{1}{\phi}|\beta|\right)}{\sqrt{2\pi\sigma^2_{\hat{\beta}}}\left[\exp\left(\frac{\bar{\beta}^2_+}{2\sigma^2_{\hat{\beta}}}\right)\left(1-\Phi\left(-\frac{\bar{\beta}_+}{\sigma_{\hat{\beta}}}\right)\right) + \exp\left(\frac{\bar{\beta}^2_-}{2\sigma^2_{\hat{\beta}}}\right)\Phi\left(-\frac{\bar{\beta}_-}{\sigma_{\hat{\beta}}}\right)\right]}$$

where all terms are defined in either model (5.2.1) *or Theorem 5.1.*

Proof. Start with Bayes' law

$$\begin{aligned} f(\beta|\boldsymbol{y}) &= \frac{f(\boldsymbol{y}|\beta)f(\beta)}{f(\boldsymbol{y})} \\ &= \frac{1}{f(\boldsymbol{y})2\phi(2\pi\sigma^2)^{\frac{n}{2}}}\exp\left(-\frac{1}{2\sigma^2}\boldsymbol{y}^T\boldsymbol{y} + \frac{1}{\sigma^2}\boldsymbol{y}^T\boldsymbol{l}\beta - \frac{1}{2\sigma^2}\beta^2 - \frac{1}{\phi}|\beta|\right). \end{aligned}$$

Substitute in the expression for $f(\boldsymbol{y})$ from Theorem 5.1 and notice that $\sigma^2_{\hat{\beta}} = \sigma^2(\boldsymbol{l}^T\boldsymbol{l})^{-1}$. □

This distribution is non-smooth at $\beta = 0$ due to the presence of the absolute value function. The distribution for small values of ϕ resembles a spike at zero. For larger values of ϕ it resembles a truncated normal where the non-truncated part has not been re-scaled. For even larger values of ϕ the predictive distribution represents a normal distribution. The predictive distribution is plotted against increasing values of ϕ in Figure 5.1. Also plotted in Figure 5.1 are the mean and mode of the predictive distribution. The expression for the mean is derived in the next section and the mode is given in (4.7.2). For small values of ϕ the mode is zero while the mean is small but non-zero. As ϕ is increased the mean increases, as does the mode but at a differing rate. For large ϕ the mean and mode tend to coincide.

Cumulative Predictive Distribution

For inferential purposes is often useful to make probability statements about the random effect β. Such statements take the form $\mathrm{P}(\beta < B|\boldsymbol{y})$ or $\mathrm{P}(\beta > B|\boldsymbol{y}) = 1 - \mathrm{P}(\beta \leq B|\boldsymbol{y})$ for some value B. These both involve the cumulative distribution function (CDF) for the predictive distribution. In the single random effect situation it is possible to express this function explicitly.

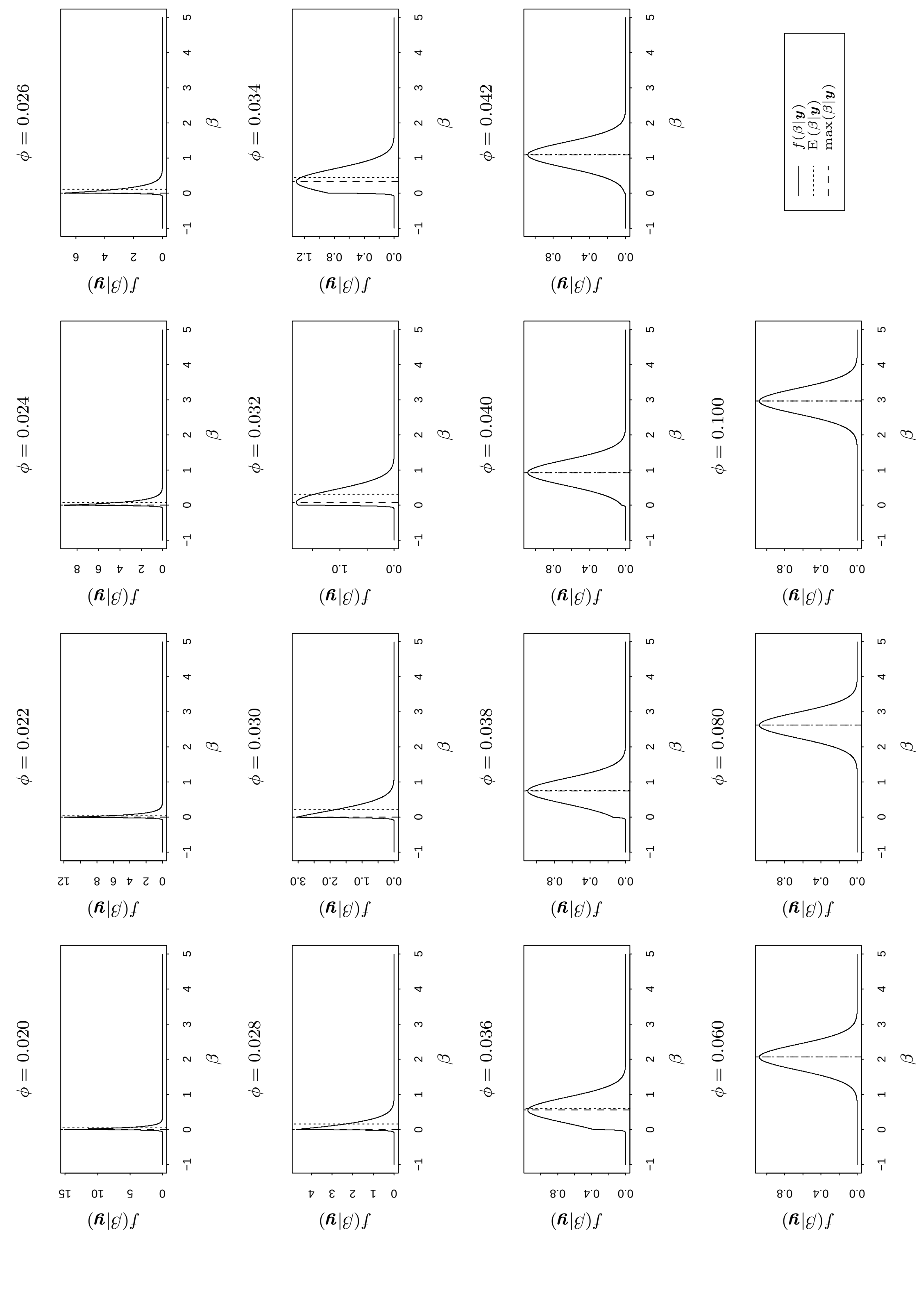

Figure 5.1: Predictive densities for varying ϕ. Common values across all plots are $\hat{\beta} = 4.33$, $\hat{\sigma}^2 = 9.02$ and $\operatorname{var}\left(\hat{\beta}\right) = 0.16$.

Theorem 5.3. *The CDF for the predictive distribution for model* (5.2.1) *is given by*

$$\mathrm{P}(\beta < B|\boldsymbol{y}) = \begin{cases} \frac{1}{d}\exp\left(\frac{\bar{\beta}_-^2}{2\sigma_{\hat{\beta}}^2}\right)\Phi\left(\frac{B-\bar{\beta}_-}{\sigma_{\hat{\beta}}}\right) & \text{for } B \leq 0 \\ \frac{1}{d}\left[\exp\left(\frac{\bar{\beta}_+^2}{2\sigma_{\hat{\beta}}^2}\right)\left(\Phi\left(\frac{B-\bar{\beta}_+}{\sigma_{\hat{\beta}}}\right) - \Phi\left(-\frac{\bar{\beta}_+}{\sigma_{\hat{\beta}}}\right)\right)\right. & \\ \qquad\left. + \exp\left(\frac{\bar{\beta}_-^2}{2\sigma_{\hat{\beta}}^2}\right)\Phi\left(-\frac{\bar{\beta}_-}{\sigma_{\hat{\beta}}}\right)\right] & \text{for } B > 0 \end{cases}$$

where

$$d = \exp\left(\frac{\bar{\beta}_+^2}{2\sigma_{\hat{\beta}}^2}\right)\left(1 - \Phi\left(-\frac{\bar{\beta}_+}{\sigma_{\hat{\beta}}}\right)\right) + \exp\left(\frac{\bar{\beta}_-^2}{2\sigma_{\hat{\beta}}^2}\right)\Phi\left(-\frac{\bar{\beta}_-}{\sigma_{\hat{\beta}}}\right).$$

Proof. The proof will only be given for $B > 0$. The proof for $B \leq 0$ follows similarly. The CDF is given by

$$\mathrm{P}(\beta \leq B|\boldsymbol{y}) = \int_{-\infty}^{0} f(\beta|\boldsymbol{y})d\beta + \int_{0}^{B} f(\beta|\boldsymbol{y})d\beta$$

From Theorem 5.2 and using the methods in the proof of Theorem 5.1 gives

$$\mathrm{P}(\beta \leq B|\boldsymbol{y}) = \frac{1}{d}\Phi\left(-\frac{\bar{\beta}_-}{\sigma_{\hat{\beta}}}\right) + \frac{1}{d\sqrt{2\pi\sigma_{\hat{\beta}}^2}}\exp\left(\frac{\bar{\beta}_+^2}{2\sigma_{\hat{\beta}}^2}\right)\int_0^B \exp\left(-\frac{1}{2\sigma_{\hat{\beta}}^2}(\beta - \bar{\beta}_+)^2\right)d\beta,$$

using the change of variable described in the proof of Theorem 5.1

$$= \frac{1}{d}\Phi\left(-\frac{\bar{\beta}_-}{\sigma_{\hat{\beta}}}\right) + \frac{1}{d\sqrt{2\pi\sigma_{\hat{\beta}}^2}}\exp\left(\frac{\beta_+^2}{2\sigma_{\hat{\beta}}^2}\right)\int_{-\frac{\bar{\beta}_+}{\sigma_{\hat{\beta}}}}^{\frac{B-\bar{\beta}_+}{\sigma_{\hat{\beta}}}} \sigma_{\hat{\beta}}\exp\left(-\frac{1}{2}u^2\right)du$$

$$= \frac{1}{d}\Phi\left(-\frac{\bar{\beta}_-}{\sigma_{\hat{\beta}}}\right) + \frac{1}{d}\exp\left(\frac{\beta_+^2}{2\sigma_{\hat{\beta}}^2}\right)\left(\Phi\left(\frac{B-\bar{\beta}_+}{\sigma_{\hat{\beta}}}\right) - \Phi\left(-\frac{\bar{\beta}_+}{\sigma_{\hat{\beta}}}\right)\right).$$

□

Typically of special interest are the probabilities $\mathrm{P}(\beta < 0|\boldsymbol{y})$ and $\mathrm{P}(\beta > 0|\boldsymbol{y})$. If one of the quantities $\bar{\beta}_+$ or $\bar{\beta}_-$ is much larger than the other then the CDF given in Theorem 5.3 has a remarkably simple form. For illustration, suppose that the LASSO estimate is large and negative. Then

$$d \approx \exp\left(\frac{\bar{\beta}_-^2}{2\sigma_{\hat{\beta}}^2}\right)\Phi\left(-\frac{\bar{\beta}_-}{\sigma_{\hat{\beta}}}\right)$$

and $\mathrm{P}(\beta < 0|\boldsymbol{y}) \approx 1$. The opposite is true if the LASSO estimate is large and positive. This simple observation shows that the probability of an effect being less than (or greater than) zero depends mostly on the size of the LASSO estimate w.r.t. the standard deviation of the least squares estimate, $\sigma_{\hat{\beta}}$.

Expected Value of Predictive Distribution

Using a similar method to that used for the marginal distribution, the expected value of the predictive distribution can be evaluated.

Theorem 5.4. *The expected value of the predictive distribution $f(\beta|\boldsymbol{y})$ is*

$$\mathrm{E}\left(\beta|\boldsymbol{y}\right) = \hat{\beta} + \frac{\sigma^2_{\hat{\beta}}}{\phi}\left[\frac{\exp\left(\frac{\hat{\beta}}{\phi}\right)\Phi\left(-\frac{\bar{\beta}_-}{\sigma_{\hat{\beta}}}\right) - \exp\left(-\frac{\hat{\beta}}{\phi}\right)\left(1-\Phi\left(-\frac{\bar{\beta}_+}{\sigma_{\hat{\beta}}}\right)\right)}{\exp\left(\frac{\hat{\beta}}{\phi}\right)\Phi\left(-\frac{\bar{\beta}_-}{\sigma_{\hat{\beta}}}\right) + \exp\left(-\frac{\hat{\beta}}{\phi}\right)\left(1-\Phi\left(-\frac{\bar{\beta}_+}{\sigma_{\hat{\beta}}}\right)\right)}\right]$$

where all parameters are defined in Theorem 5.1 or in model (5.2.1)*.*

Proof.

$$\begin{aligned}\mathrm{E}\left(\beta|\boldsymbol{y}\right) &= \int_{-\infty}^{\infty}\beta f(\beta|\boldsymbol{y})d\beta \\ &= \frac{1}{f(\boldsymbol{y})}\left(\int_0^{\infty}\beta f(\boldsymbol{y}|\beta)f(\beta)d\beta + \int_{-\infty}^{0}\beta f(\boldsymbol{y}|\beta)f(\beta)d\beta\right) \\ &= \frac{1}{f(\boldsymbol{y})}(A+B) \qquad\qquad \text{say.}\end{aligned}$$

First consider the integral A, using an argument similar to that used in the proof of Theorem 5.1.

$$\begin{aligned}A &= \frac{1}{2\phi(2\pi\sigma^2)^{\frac{n}{2}}}\exp\left(-\frac{1}{2\sigma^2}\boldsymbol{y}^T\boldsymbol{y}\right)\int_0^{\infty}\beta\exp\left(-\frac{1}{2\sigma^2_{\hat{\beta}}}(\beta^2-2\beta\bar{\beta}_+)\right)d\beta \\ &= \frac{1}{2\phi(2\pi\sigma^2)^{\frac{n}{2}}}\exp\left(-\frac{1}{2\sigma^2}\boldsymbol{y}^T\boldsymbol{y}\right)\left[\int_0^{\infty}(\beta-\bar{\beta}_+)\exp\left(-\frac{1}{2\sigma^2_{\hat{\beta}}}(\beta^2-2\beta\bar{\beta}_+)\right)d\beta\right. \\ &\qquad\qquad\left. + \bar{\beta}_+\int_0^{\infty}\exp\left(-\frac{1}{2\sigma^2_{\hat{\beta}}}(\beta^2-2\beta\bar{\beta}_+)\right)d\beta\right] \\ &= \frac{1}{2\phi(2\pi\sigma^2)^{\frac{n}{2}}}\exp\left(-\frac{1}{2\sigma^2}\boldsymbol{y}^T\boldsymbol{y}\right)\left[\int_0^{\infty}(\beta-\bar{\beta}_+)\exp\left(-\frac{1}{2\sigma^2_{\hat{\beta}}}(\beta^2-2\beta\bar{\beta}_+)\right)d\beta\right. \\ &\qquad\qquad\left. + \sqrt{2\pi\sigma^2_{\hat{\beta}}}\bar{\beta}_+\exp\left(\frac{\bar{\beta}^2_+}{2\sigma^2_{\hat{\beta}}}\right)\left(1-\Phi\left(-\frac{\bar{\beta}_+}{\sigma_{\hat{\beta}}}\right)\right)\right].\end{aligned}$$

The first integral is simplified by a change of variables. Let

$$u = \frac{\beta^2-2\beta\bar{\beta}_+}{2\sigma^2_{\hat{\beta}}} \qquad \text{giving} \qquad \frac{du}{d\beta} = \frac{\beta-\bar{\beta}_+}{\sigma^2_{\hat{\beta}}}.$$

This substitution gives

$$\begin{aligned}\int_0^{\infty}(\beta-\bar{\beta}_+)\exp\left(-\frac{1}{2\sigma^2_{\hat{\beta}}}(\beta^2-2\beta\bar{\beta}_+)\right)d\beta &= \sigma^2_{\hat{\beta}}\int_0^{\infty}\exp\left(-u\right)\frac{du}{d\beta}d\beta \\ &= \sigma^2_{\hat{\beta}}.\end{aligned}$$

The last equality holds as the integrand is the exponential distribution. The integral A can be expressed as

$$\begin{aligned}
A &= \frac{1}{2\phi(2\pi\sigma^2)^{\frac{n}{2}}}\exp\left(-\frac{1}{2\sigma^2}\boldsymbol{y}^T\boldsymbol{y}\right)\left[\sigma_{\hat{\beta}}^2 + \sqrt{2\pi\sigma_{\hat{\beta}}^2}\bar{\beta}_+\exp\left(\frac{\bar{\beta}_+^2}{2\sigma_{\hat{\beta}}^2}\right)\left(1-\Phi\left(-\frac{\bar{\beta}_+}{\sigma_{\hat{\beta}}}\right)\right)\right] \\
&= \frac{1}{2\phi(2\pi\sigma^2)^{\frac{n}{2}}}\exp\left(-\frac{1}{2\sigma^2}\boldsymbol{y}^T\boldsymbol{y}\right)\times \\
&\qquad \left[\sigma_{\hat{\beta}}^2 + \sqrt{2\pi\sigma_{\hat{\beta}}^2}\bar{\beta}_+\exp\left(\frac{\hat{\beta}^2}{2\sigma_{\hat{\beta}}^2}+\frac{\sigma_{\hat{\beta}}^2}{2\phi^2}\right)\exp\left(-\frac{\hat{\beta}}{\phi}\right)\left(1-\Phi\left(-\frac{\bar{\beta}_+}{\sigma_{\hat{\beta}}}\right)\right)\right].
\end{aligned}$$

A similar sequence of operations gives the integral B

$$\begin{aligned}
B &= \frac{1}{2\phi(2\pi\sigma^2)^{\frac{n}{2}}}\exp\left(-\frac{1}{2\sigma^2}\boldsymbol{y}^T\boldsymbol{y}\right)\times \\
&\qquad \left[-\sigma_{\hat{\beta}}^2 + \sqrt{2\pi\sigma_{\hat{\beta}}^2}\bar{\beta}_-\exp\left(\frac{\hat{\beta}^2}{2\sigma_{\hat{\beta}}^2}+\frac{\sigma_{\hat{\beta}}^2}{2\phi^2}\right)\exp\left(\frac{\hat{\beta}}{\phi}\right)\Phi\left(-\frac{\bar{\beta}_-}{\sigma_{\hat{\beta}}}\right)\right].
\end{aligned}$$

Using the form of $f(\boldsymbol{y})$ given in Corollary 5.1 the expected value is

$$\begin{aligned}
\mathrm{E}\left(\beta|\boldsymbol{y}\right) &= \frac{B+A}{f(\boldsymbol{y})} \\
&= \frac{\bar{\beta}_-\exp\left(\frac{\hat{\beta}}{\phi}\right)\Phi\left(-\frac{\bar{\beta}_-}{\sigma_{\hat{\beta}}}\right)+\bar{\beta}_+\exp\left(-\frac{\hat{\beta}}{\phi}\right)\left(1-\Phi\left(-\frac{\bar{\beta}_+}{\sigma_{\hat{\beta}}}\right)\right)}{\exp\left(\frac{\hat{\beta}}{\phi}\right)\Phi\left(-\frac{\bar{\beta}_-}{\sigma_{\hat{\beta}}}\right)+\exp\left(-\frac{\hat{\beta}}{\phi}\right)\left(1-\Phi\left(-\frac{\bar{\beta}_+}{\sigma_{\hat{\beta}}}\right)\right)} \\
&= \frac{\left(\hat{\beta}+\frac{\sigma_{\hat{\beta}}^2}{\phi}\right)\exp\left(\frac{\hat{\beta}}{\phi}\right)\Phi\left(-\frac{\bar{\beta}_-}{\sigma_{\hat{\beta}}}\right)+\left(\hat{\beta}-\frac{\sigma_{\hat{\beta}}^2}{\phi}\right)\exp\left(-\frac{\hat{\beta}}{\phi}\right)\left(1-\Phi\left(-\frac{\bar{\beta}_+}{\sigma_{\hat{\beta}}}\right)\right)}{\exp\left(\frac{\hat{\beta}}{\phi}\right)\Phi\left(-\frac{\bar{\beta}_-}{\sigma_{\hat{\beta}}}\right)+\exp\left(-\frac{\hat{\beta}}{\phi}\right)\left(1-\Phi\left(-\frac{\bar{\beta}_+}{\sigma_{\hat{\beta}}}\right)\right)} \\
&= \hat{\beta}+\frac{\sigma_{\hat{\beta}}^2}{\phi}\left[\frac{\exp\left(\frac{\hat{\beta}}{\phi}\right)\Phi\left(-\frac{\bar{\beta}_-}{\sigma_{\hat{\beta}}}\right)-\exp\left(-\frac{\hat{\beta}}{\phi}\right)\left(1-\Phi\left(-\frac{\bar{\beta}_+}{\sigma_{\hat{\beta}}}\right)\right)}{\exp\left(\frac{\hat{\beta}}{\phi}\right)\Phi\left(-\frac{\bar{\beta}_-}{\sigma_{\hat{\beta}}}\right)+\exp\left(-\frac{\hat{\beta}}{\phi}\right)\left(1-\Phi\left(-\frac{\bar{\beta}_+}{\sigma_{\hat{\beta}}}\right)\right)}\right].
\end{aligned}$$

□

The expected value in Theorem 5.4, like the ridge regression and the LASSO, is a shrinkage estimate. That is, the estimate is moved towards zero from the least squares estimate. The amount of shrinkage is governed by the additive shrinkage term in Theorem 5.4. The type of shrinkage adjustment in the expected value in Theorem 5.4 and the LASSO is additive in contrast to the ridge regression estimate which is multiplicative.

The second part of this additive shrinkage term is, theoretically, in the interval $(-1,1)$. The normal cumulative distribution function does not have a closed form and hence, in practice, must be evaluated numerically. For small ϕ this causes the computed estimate to be unstable as algorithms to perform this numerical evaluation are not extremely accurate in the tails of the normal distribution.

If the dispersion parameter of the random effect, ϕ, is allowed to tend to infinity, the mean estimate will tend towards the least squares estimate $\hat{\beta}$. This arises as the denominator in the additive shrinkage term will cause the adjustment to tend to zero.

Comparison of Mean and Mode of the Predictive Distribution

The mode of the predictive distribution (the LASSO estimate) is the same as the mode of the joint distribution.

$$\begin{aligned}\text{mode}\,(f(\beta|\boldsymbol{y})) &= \bar{\beta} \\ &= (\boldsymbol{l}^T\boldsymbol{l})^{-1}\boldsymbol{l}^T\boldsymbol{y} - \frac{\sigma^2}{\phi}(\boldsymbol{l}^T\boldsymbol{l})^{-1}\bar{v} \\ &= \hat{\beta} - \frac{\sigma^2_{\hat{\beta}}}{\phi}\bar{v}\end{aligned}$$

where $\bar{v} = \text{sign}(\bar{\beta})$ or equivalently for the scalar effect $\text{sign}(\hat{\beta})$.

The functional form of the mean and mode are strikingly similar. The difference is in the second, or shrinkage adjustment, term. The adjustment in the mean is based on a probability argument. If the LASSO estimate is either very large and positive or negative, then this term will approximate the sign of the LASSO estimate.

The mean and mode will converge to the least squares estimate as ϕ increases. The mean estimate will always be larger in absolute value than the LASSO estimate. For a non-zero mode, the probability adjustment term for the mean estimate tends to ±1 whereas the adjustment in the LASSO estimate is identically ±1. For a zero mode the mean estimate is non-zero and hence will be larger in absolute value. The mean and LASSO estimates are plotted against ϕ in Figure 5.2 demonstrating the points raised.

Variance of Predictive Distribution

The variance of the predictive distribution can be important for making probability statements about the random effects. In the simple model (5.2.1) the predictive distribution variance is far from trivial. It can be evaluated analytically but it does not provide great insight into the nature of the distribution. Nevertheless, it is now given for completeness.

Theorem 5.5. *The variance of the predictive distribution $f(\beta|\boldsymbol{y})$ for scalar β is*

$$\text{var}\,(\beta|\boldsymbol{y}) = \sigma^2_{\hat{\beta}} - \frac{2}{\sqrt{2\pi\sigma^2_{\hat{\beta}}}}\frac{\sigma^4_{\hat{\beta}}}{\phi}\frac{1}{d} - 4\hat{\beta}\frac{\sigma^2_{\hat{\beta}}}{\phi}\frac{a}{d} + \frac{\sigma^4_{\hat{\beta}}}{\phi^2}\left(1 - \frac{a^2}{d^2}\right) \qquad \textit{where}$$

$$a = \exp\left(\frac{\bar{\beta}}{\phi}\right)\Phi\left(-\frac{\bar{\beta}_-}{\sigma_{\hat{\beta}}}\right) - \exp\left(-\frac{\bar{\beta}}{\phi}\right)\left(1 - \Phi\left(-\frac{\bar{\beta}_+}{\sigma_{\hat{\beta}}}\right)\right)$$

$$d = \exp\left(\frac{\bar{\beta}}{\phi}\right)\Phi\left(-\frac{\bar{\beta}_-}{\sigma_{\hat{\beta}}}\right) + \exp\left(-\frac{\bar{\beta}}{\phi}\right)\left(1 - \Phi\left(-\frac{\bar{\beta}_+}{\sigma_{\hat{\beta}}}\right)\right)$$

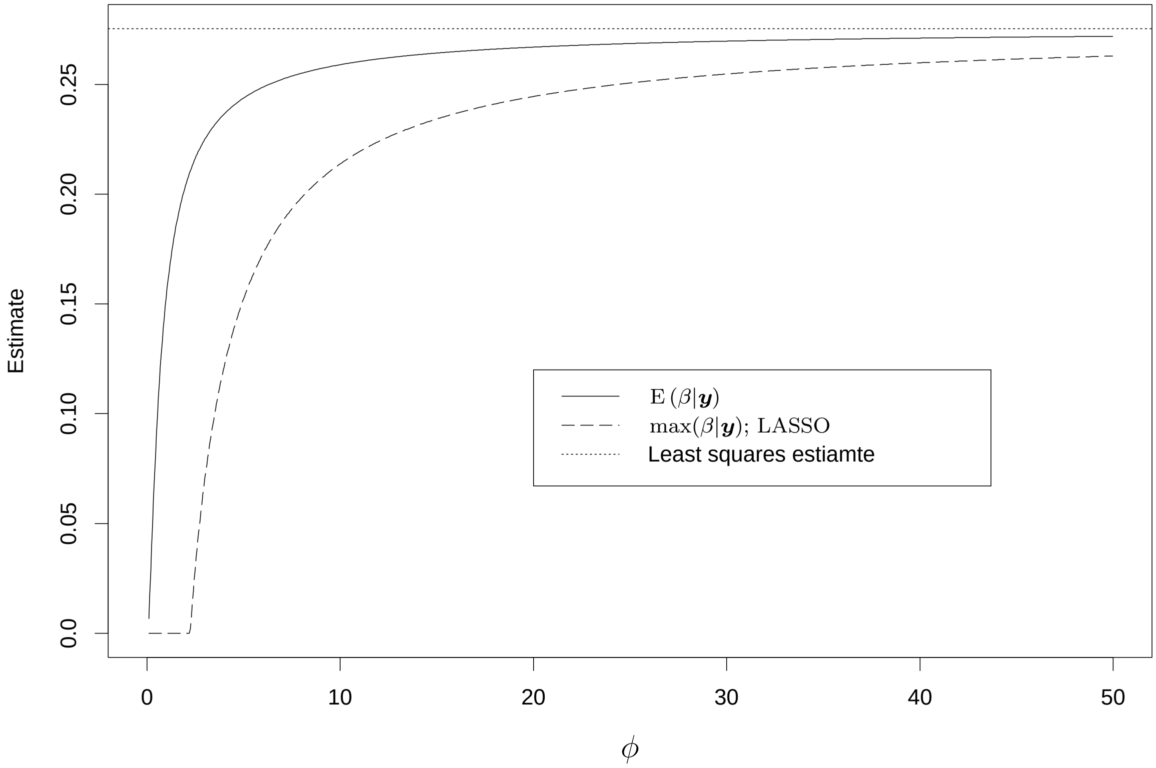

Figure 5.2: Plot of the least squares estimate, the predictive distribution's mean and mode (LASSO estimate) as a function of ϕ. $\hat{\boldsymbol{\beta}} = 0.275$, $\hat{\sigma^2} = 92.22$.

and all other parameters are defined in Theorem 5.1 or in model (5.2.1).

Proof. The variance is given by

$$\mathrm{var}(\beta|\boldsymbol{y}) = \exp\left(\beta^2|\boldsymbol{y}\right) - \left(\mathrm{E}(\beta|\boldsymbol{y})\right)^2.$$

The second term is available immediately from Theorem 5.4. Note that

$$\mathrm{E}(\beta|\boldsymbol{y}) = \hat{\beta} + \frac{\sigma^2_{\hat{\beta}}}{\phi}\frac{a}{d}$$

so that

$$\left(\mathrm{E}(\beta|\boldsymbol{y})\right)^2 = \hat{\beta}^2 - 2\hat{\beta}\frac{\sigma^2_{\hat{\beta}}}{\phi}\frac{a}{d} + \frac{\sigma^4_{\hat{\beta}}}{\phi^2}\frac{a^2}{d^2}$$

The first term can be expressed as (see proofs of Theorems 5.1 and 5.4)

$$\begin{aligned}
\mathrm{E}\left(\beta^2|\boldsymbol{y}\right) &= \frac{1}{f(\boldsymbol{y})}\int_{-\infty}^{\infty}\beta^2 f(\boldsymbol{y}|\beta)f(\beta)d\beta \\
&= \frac{1}{f(\boldsymbol{y})}\frac{1}{2\phi(2\pi\sigma^2)^{\frac{n}{2}}}\mathrm{exp}\left(-\frac{1}{2\sigma^2}\boldsymbol{y}^T\boldsymbol{y}\right)\times \\
&\quad \left[\int_0^{\infty}\beta^2\mathrm{exp}\left(-\frac{1}{2\sigma^2_{\hat{\beta}}}\left(\beta^2-2\beta\bar{\beta}_+\right)\right)d\beta + \int_{-\infty}^{0}\beta^2\mathrm{exp}\left(-\frac{1}{2\sigma^2_{\hat{\beta}}}\left(\beta^2-2\beta\bar{\beta}_-\right)\right)\right] \\
&= \frac{1}{f(\boldsymbol{y})}\frac{1}{2\phi(2\pi\sigma^2)^{\frac{n}{2}}}\mathrm{exp}\left(-\frac{1}{2\sigma^2}\boldsymbol{y}^T\boldsymbol{y}\right)[A+B] \qquad \text{say}
\end{aligned}$$

$$\begin{aligned}
A &= \mathrm{exp}\left(\frac{\bar{\beta}_+^2}{2\sigma^2_{\hat{\beta}}}\right)\int_0^{\infty}(\beta-\bar{\beta}_+)^2\mathrm{exp}\left(-\frac{1}{2\sigma^2_{\hat{\beta}}}\left(\beta-\bar{\beta}_+\right)^2\right)d\beta \\
&\quad + 2\bar{\beta}_+\int_0^{\infty}\mathrm{exp}\left(-\frac{1}{2\sigma^2_{\hat{\beta}}}\left(\beta^2-2\beta\bar{\beta}_+\right)\right)d\beta \\
&\quad - \mathrm{exp}\left(\frac{\bar{\beta}_+^2}{2\sigma^2_{\hat{\beta}}}\right)\int_0^{\infty}\mathrm{exp}\left(-\frac{1}{2\sigma^2_{\hat{\beta}}}\left(\beta-\bar{\beta}_+\right)^2\right)d\beta \\
&= A_1 + A_2 - A_3 \qquad \text{say.}
\end{aligned}$$

From the proof of Theorem 5.4

$$A_2 = 2\sigma^2_{\hat{\beta}}\bar{\beta}_+ + 2\sqrt{2\pi\sigma^2_{\hat{\beta}}}\bar{\beta}_+^2\mathrm{exp}\left(\frac{\bar{\beta}_+^2}{2\sigma^2_{\hat{\beta}}}\right)\left(1-\Phi\left(-\frac{\bar{\beta}_+}{\sigma_{\hat{\beta}}}\right)\right).$$

From the proof of Theorem 5.1

$$A_3 = \sqrt{2\pi\sigma^2_{\hat{\beta}}}\bar{\beta}_+^2\mathrm{exp}\left(\frac{\bar{\beta}_+^2}{2\sigma^2_{\hat{\beta}}}\right)\left(1-\Phi\left(-\frac{\bar{\beta}_+}{\sigma_{\hat{\beta}}}\right)\right).$$

These two equations give

$$A_2 - A_3 = 2\sigma^2_{\hat{\beta}}\bar{\beta}_+ + \sqrt{2\pi\sigma^2_{\hat{\beta}}}\bar{\beta}_+^2\mathrm{exp}\left(\frac{\bar{\beta}_+^2}{2\sigma^2_{\hat{\beta}}}\right)\left(1-\Phi\left(-\frac{\bar{\beta}_+}{\sigma_{\hat{\beta}}}\right)\right).$$

The integral in A_1 is evaluated using integration by parts (e.g. Apostol, 1967). Let

$$\begin{aligned}
f(\beta) &= \beta-\bar{\beta}_+ & \Rightarrow \quad & f'(\beta) = 1 \qquad \text{and} \\
g'(\beta) &= (\beta-\bar{\beta}_+)\mathrm{exp}\left(-\frac{1}{2\sigma^2_{\hat{\beta}}}(\beta-\bar{\beta}_+)^2\right) & \Rightarrow \quad & g(\beta) = -\sigma^2_{\hat{\beta}}\mathrm{exp}\left(-\frac{1}{2\sigma^2_{\hat{\beta}}}(\beta-\bar{\beta}_+)^2\right)
\end{aligned}$$

then

$$\begin{aligned}
\exp\left(-\frac{\bar{\beta}_+^2}{2\sigma_{\hat{\beta}}^2}\right) A_1 &= f(\beta)g(\beta)\Big|_0^\infty - \int_0^\infty f'(\beta)g(\beta)d\beta \\
&= \left[-\beta\sigma_{\hat{\beta}}^2 \exp\left(-\frac{1}{2\sigma_{\hat{\beta}}^2}(\beta-\bar{\beta}_+)^2\right) + \bar{\beta}_+\sigma_{\hat{\beta}}^2 \exp\left(-\frac{1}{2\sigma_{\hat{\beta}}^2}(\beta-\bar{\beta})^2\right)\right]_0^\infty \\
&\qquad + \sigma_{\hat{\beta}}^2\sqrt{2\pi\sigma_{\hat{\beta}}^2}\left(1-\Phi\left(-\frac{\bar{\beta}_+}{\sigma_{\hat{\beta}}}\right)\right).
\end{aligned}$$

Appealing to L'Hôpitals rule for the value of $f(\beta)g(\beta)$ evaluated at ∞ yields

$$A_1 = -\bar{\beta}_+\sigma_{\hat{\beta}}^2 + \sigma_{\hat{\beta}}^2\sqrt{2\pi\sigma_{\hat{\beta}}^2}\exp\left(\frac{\bar{\beta}_+^2}{2\sigma_{\hat{\beta}}^2}\right)\left(1-\Phi\left(-\frac{\bar{\beta}_+}{\sigma_{\hat{\beta}}}\right)\right)$$

and

$$A = \sigma_{\hat{\beta}}^2\bar{\beta}_+ + \sqrt{2\pi\sigma_{\hat{\beta}}^2}\exp\left(\frac{\bar{\beta}_+^2}{2\sigma_{\hat{\beta}}^2}\right)\left(1-\Phi\left(-\frac{\bar{\beta}_+}{\sigma_{\hat{\beta}}}\right)\right)\left(\bar{\beta}_+^2+\sigma_{\hat{\beta}}^2\right).$$

An analogous sequence of operations gives

$$B = -\bar{\beta}_-\sigma_{\hat{\beta}}^2 + \sqrt{2\pi\sigma_{\hat{\beta}}^2}\exp\left(\frac{\bar{\beta}_-^2}{2\sigma_{\hat{\beta}}^2}\right)\Phi\left(-\frac{\bar{\beta}_-}{\sigma_{\hat{\beta}}}\right)\left(\bar{\beta}_-^2+\sigma_{\hat{\beta}}^2\right).$$

The part of $\mathrm{E}\left(\beta^2|\boldsymbol{y}\right)$ involving the integrals can now be given

$$A+B = -2\frac{\sigma_{\hat{\beta}}^4}{\phi^2} + \sqrt{2\pi\sigma_{\hat{\beta}}^2}\hat{\beta}^2 d - 2\sqrt{2\pi\sigma_{\hat{\beta}}^2}\hat{\beta}\frac{\sigma_{\hat{\beta}}^2}{\phi}a + \sqrt{2\pi\sigma_{\hat{\beta}}^2}\frac{\sigma_{\hat{\beta}}^4}{\phi^2}d + \sqrt{2\pi\sigma_{\hat{\beta}}^2}\sigma_{\hat{\beta}}^2 d$$

as can $\mathrm{E}\left(\beta^2|\boldsymbol{y}\right)$

$$\begin{aligned}
\mathrm{E}\left(\beta^2|\boldsymbol{y}\right) &= \frac{1}{f(\boldsymbol{y})}\frac{1}{2\phi(2\pi\sigma^2)^{\frac{n}{2}}}\exp\left(-\frac{1}{2\sigma^2}\boldsymbol{y}^T\boldsymbol{y}\right)(A+B) \\
&= \frac{1}{\sqrt{2\pi\sigma_{\hat{\beta}}^2}}\frac{1}{d}(A+B) \\
&= -\frac{2\sigma_{\hat{\beta}}^4}{d\phi\sqrt{2\pi\sigma_{\hat{\beta}}^2}} + \hat{\beta}^2 - 2\hat{\beta}\frac{\sigma_{\hat{\beta}}^2}{\phi}\frac{a}{d} + \frac{\sigma_{\hat{\beta}}^4}{\phi^2} + \sigma_{\hat{\beta}}^2.
\end{aligned}$$

This gives the variance to be

$$\mathrm{var}\left(\beta|\boldsymbol{y}\right) = \sigma_{\hat{\beta}}^2 - \frac{2}{\sqrt{2\pi\sigma_{\hat{\beta}}^2}}\frac{\sigma_{\hat{\beta}}^4}{\phi}\frac{1}{d} - 4\hat{\beta}\frac{\sigma_{\hat{\beta}}^2}{\phi}\frac{a}{d} + \frac{\sigma_{\hat{\beta}}^4}{\phi^2}\left(1-\frac{a^2}{d^2}\right).$$

□

The only easily obtainable observation regarding this variance is that when ϕ tends to infinity, the predictive distribution's variance tends to the least squares estimate's variance $\sigma_{\hat{\beta}}^2$. From the plots of the predictive distribution in Figure 5.1, it is expected that the predictive distribution's variance should decrease as ϕ gets smaller. However, this is not obvious from the variance in Theorem 5.2.2.

5.3 Vector of Effects

A multiple dimension generalisation of the marginal distribution for a single random effect β is available. However, the expected value and variance of the predictive distribution are not easily obtained.

Consider the model containing q random effects $\boldsymbol{\beta}$

$$\boldsymbol{y} = \boldsymbol{L}\boldsymbol{\beta} + \boldsymbol{e} \tag{5.3.1}$$

where $\boldsymbol{y}$ is a $n \times 1$ vector of observed outcomes, $\boldsymbol{L}$ is a $n \times q$ design matrix, $\boldsymbol{\beta}$ is a $q \times 1$ vector of independent random effects with $\beta_i \sim \mathrm{DE}(0, 2\phi^2)$ and $\boldsymbol{e}$ is the $n \times 1$ vector of independent normal residuals with common variance σ^2.

5.3.1 *Marginal Distribution of Observations*

Theorem 5.6. *The marginal distribution for the multiple random effects model is*

$$\begin{aligned} f(\boldsymbol{y}) = {} & \frac{(2\pi)^{\frac{q}{2}}}{(2\phi)^q(2\pi\sigma^2)^{\frac{n}{2}}}|\Sigma|^{\frac{1}{2}}\exp\left(-\frac{1}{2\sigma^2}\boldsymbol{y}^T\boldsymbol{y}\right) \times \\ & \sum_{j=1}^{2^q}\left[\exp\left(\bar{\boldsymbol{\beta}}_j^T\Sigma^{-1}\bar{\boldsymbol{\beta}}_j\right)\int_{\mathcal{A}_j}\frac{1}{(2\pi)^{\frac{q}{2}}}|\Sigma|^{-\frac{1}{2}}\exp\left(-\frac{1}{2}(\boldsymbol{\beta}-\bar{\boldsymbol{\beta}}_j)^T\Sigma^{-1}(\boldsymbol{\beta}-\bar{\boldsymbol{\beta}}_j)\right)d\boldsymbol{\beta}\right] \end{aligned} \tag{5.3.2}$$

where $\{\mathcal{A}_j\}_{j=1}^{2^q}$ *is the complete set of orthants in* $\mathbb{R}^q$, $\bar{\boldsymbol{\beta}}_j = \hat{\boldsymbol{\beta}} - \frac{\sigma^2}{\phi}(\boldsymbol{L}^T\boldsymbol{L})^{-}\boldsymbol{v}_j$, $\hat{\boldsymbol{\beta}} = (\boldsymbol{L}^T\boldsymbol{L})^{-1}\boldsymbol{L}^T\boldsymbol{y}$, $\Sigma^{-1} = \frac{1}{\sigma^2}(\boldsymbol{L}^T\boldsymbol{L})$ *and* $\boldsymbol{v}_j$ *is a vector with elements* ± 1 *corresponding to the sign of the points in orthant* $\mathcal{A}_j$.

Proof.

$$\begin{aligned} f(\boldsymbol{y}) &= \int_{\mathbb{R}^q} f(\boldsymbol{y}|\boldsymbol{\beta})f(\boldsymbol{\beta}) \\ &= \frac{1}{(2\phi)^q(2\pi\sigma^2)^{\frac{n}{2}}}\int_{\mathbb{R}^q}\exp\left(-\frac{1}{2\sigma^2}(\boldsymbol{y}-\boldsymbol{L}\boldsymbol{\beta})^T(\boldsymbol{y}-\boldsymbol{L}\boldsymbol{\beta}) - \frac{1}{\phi}\|\boldsymbol{\beta}\|_1\right)d\boldsymbol{\beta} \\ &= \frac{1}{(2\phi)^q(2\pi\sigma^2)^{\frac{n}{2}}}\exp\left(-\frac{1}{2\sigma^2}\boldsymbol{y}^T\boldsymbol{y}\right) \times \\ &\qquad \int_{\mathbb{R}^q}\exp\left(-\frac{1}{2\sigma^2}\left(\boldsymbol{\beta}^T\boldsymbol{L}^T\boldsymbol{L}\boldsymbol{\beta} - \boldsymbol{\beta}^T\boldsymbol{L}^T\boldsymbol{y} - \boldsymbol{y}^T\boldsymbol{L}\boldsymbol{\beta} + \frac{\sigma^2}{\phi}\boldsymbol{v}^T\boldsymbol{\beta} + \frac{\sigma^2}{\phi}\boldsymbol{\beta}^T\boldsymbol{v}\right)\right)d\boldsymbol{\beta} \end{aligned}$$

where $\boldsymbol{v} = \mathrm{sign}(\boldsymbol{\beta})$. Now

$$\begin{aligned} f(\boldsymbol{y}) = {} & \frac{1}{(2\phi)^q(2\pi\sigma^2)^{\frac{n}{2}}}\exp\left(-\frac{1}{2\sigma^2}\boldsymbol{y}^T\boldsymbol{y}\right) \times \\ & \int_{\mathbb{R}^q}\exp\left(-\frac{1}{2}\left(\boldsymbol{\beta}^T\Sigma^{-1}\boldsymbol{\beta} - \boldsymbol{\beta}^T\Sigma^{-1}\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}^T\Sigma^{-1}\boldsymbol{\beta} + \frac{\sigma^2}{\phi}\boldsymbol{v}^T\boldsymbol{\beta} + \frac{\sigma^2}{\phi}\boldsymbol{\beta}^T\boldsymbol{v}\right)\right)d\boldsymbol{\beta} \end{aligned}$$

where $\hat{\boldsymbol{\beta}} = (\boldsymbol{L}^T\boldsymbol{L})^{-1}\boldsymbol{L}^T\boldsymbol{y}$ and $\Sigma^{-1} = \frac{1}{\sigma^2}(\boldsymbol{L}^T\boldsymbol{L})$. This gives

$$f(\boldsymbol{y}) = \frac{1}{(2\phi)^q(2\pi\sigma^2)^{\frac{n}{2}}}\exp\left(-\frac{1}{2\sigma^2}\boldsymbol{y}^T\boldsymbol{y}\right) \times$$
$$\int_{\mathbb{R}^q} \exp\left(-\frac{1}{2}\left(\boldsymbol{\beta}^T\Sigma^{-1}\boldsymbol{\beta} - \boldsymbol{\beta}^T\Sigma^{-1}\bar{\boldsymbol{\beta}} - \bar{\boldsymbol{\beta}}^T\Sigma^{-1}\boldsymbol{\beta}\right)\right) d\boldsymbol{\beta}$$

where $\bar{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}} - \frac{1}{\phi}\Sigma\boldsymbol{v} = \hat{\boldsymbol{\beta}} - \frac{\sigma^2}{\phi}(\boldsymbol{L}^T\boldsymbol{L})^{-}\boldsymbol{v}$. Complete the square in the exponent to obtain

$$f(\boldsymbol{y}) = \frac{1}{(2\phi)^q(2\pi\sigma^2)^{\frac{n}{2}}}\exp\left(-\frac{1}{2\sigma^2}\boldsymbol{y}^T\boldsymbol{y}\right)\exp\left(\bar{\boldsymbol{\beta}}^T\Sigma^{-1}\bar{\boldsymbol{\beta}}\right)$$
$$\int_{\mathbb{R}^q} \exp\left(-\frac{1}{2}(\boldsymbol{\beta} - \bar{\boldsymbol{\beta}})^T\Sigma^{-1}(\boldsymbol{\beta} - \bar{\boldsymbol{\beta}})\right) d\boldsymbol{\beta}.$$

The form of $\boldsymbol{v}$ will vary depending on which orthant of $\mathbb{R}^q$ is being considered. There are 2^q possible forms for $\boldsymbol{v}$. Evaluation of this integral proceeds by separating the integral into a sum of 2^q integrals, each one integrating over one of the orthants $\{\mathcal{A}\}_{j=1}^{2^q}$. The sum gives the result. □

The evaluation of a multivariate normal probability requires numerical methods. This is a well known statistical problem. Acceptable solutions have been found where the number of dimensions q is not large and when the value of the probability is moderate (preferably around 0.5), for example see Kotz et al. (2000) or Gassmann et al. (2002). Unfortunately, in this application q may be large and many, maybe all, of the probabilities will be close to zero or one. For these reasons, along with the complexity of the form for the marginal distribution, an alternative form must be obtained for the marginal distribution if it is to be used. This problem is addressed in Chapter 7 and 10 where approximate marginal distributions are developed.

5.4 Reduction of LASSO Estimates

The LASSO estimates given in Chapter 4 require a generalised inverse of the $q \times q$ matrix $\boldsymbol{L}^T\boldsymbol{L}$. This can be avoided in situations where the partition of the effects ($\boldsymbol{\beta}$) into those that will be estimated to be non-zero ($\boldsymbol{\beta}_{\mathcal{S}}$) and those that will be estimated to be zero ($\boldsymbol{\beta}_{\mathcal{Z}}$) is known. The mechanism to perform this comes from an alternative expression to those in Chapter 4 for the LASSO effects conditional on the partition.

Theorem 5.7. *Consider that the dispersion parameters, the partition of the random effects in model* (5.3.1) *into those that will be estimated to be non-zero ($\boldsymbol{\beta}_{\mathcal{S}}$) and those estimated to be zero ($\boldsymbol{\beta}_{\mathcal{Z}}$) are all known. Further, assume that $q < n$ so that L^TL is non-singular. Arrange the effects so that the non-zero effects are in the first positions and arrange $\boldsymbol{v}$ and*

$\boldsymbol{L}$ *conformably. The LASSO estimates in Chapter 4 can be given by*

$$\begin{aligned}\bar{\boldsymbol{\beta}} &= \begin{bmatrix}\bar{\boldsymbol{\beta}}_{\mathcal{S}}\\ \bar{\boldsymbol{\beta}}_{\mathcal{Z}}\end{bmatrix}\\ &= \begin{bmatrix}(\boldsymbol{L}_{\mathcal{S}}^T\boldsymbol{L}_{\mathcal{S}})^{-1}\boldsymbol{L}_{\mathcal{S}}\boldsymbol{y} - \frac{\sigma^2}{\phi}(\boldsymbol{L}_{\mathcal{S}}^T\boldsymbol{L}_{\mathcal{S}})^{-1}\boldsymbol{v}_{\mathcal{S}}\\ \boldsymbol{0}\end{bmatrix}.\end{aligned}$$

Proof. Define

$$(\boldsymbol{L}^T\boldsymbol{L})^{-1} = \begin{bmatrix}(\boldsymbol{L}^T\boldsymbol{L})^{-1}_{\mathcal{S}\mathcal{S}} & (\boldsymbol{L}^T\boldsymbol{L})^{-1}_{\mathcal{S}\mathcal{Z}}\\ (\boldsymbol{L}^T\boldsymbol{L})^{-1}_{\mathcal{Z}\mathcal{S}} & (\boldsymbol{L}^T\boldsymbol{L})^{-1}_{\mathcal{Z}\mathcal{Z}}\end{bmatrix}$$

and let $(\boldsymbol{L}^T\boldsymbol{L})$ be likewise partitioned. From Chapter 4 it is known that

$$\begin{aligned}\bar{\boldsymbol{\beta}}_{\mathcal{Z}} &= \boldsymbol{0}\\ &= (\boldsymbol{L}^T\boldsymbol{L})^{-1}_{\boldsymbol{z}\cdot}\boldsymbol{L}^T\boldsymbol{y} - \frac{\sigma^2}{\phi}(\boldsymbol{L}^T\boldsymbol{L})^{-1}_{\boldsymbol{z}\cdot}\bar{\boldsymbol{v}}\\ &= (\boldsymbol{L}^T\boldsymbol{L})^{-1}_{\mathcal{Z}\mathcal{S}}\boldsymbol{L}^T_{\mathcal{S}}\boldsymbol{y} + (\boldsymbol{L}^T\boldsymbol{L})^{-1}_{\mathcal{Z}\mathcal{Z}}\boldsymbol{L}^T_{\mathcal{Z}}\boldsymbol{y} - \frac{\sigma^2}{\phi}(\boldsymbol{L}^T\boldsymbol{L})^{-1}_{\mathcal{Z}\mathcal{S}}\bar{\boldsymbol{v}}_{\mathcal{S}} - \frac{\sigma^2}{\phi}(\boldsymbol{L}^T\boldsymbol{L})^{-1}_{\mathcal{Z}\mathcal{Z}}\bar{\boldsymbol{v}}_{\mathcal{Z}}\end{aligned}$$

giving

$$\boldsymbol{L}^T_{\mathcal{Z}}\boldsymbol{y} - \frac{\sigma^2}{\phi}\bar{\boldsymbol{v}}_{\mathcal{Z}} = -\left[(\boldsymbol{L}^T\boldsymbol{L})^{-1}_{\mathcal{Z}\mathcal{Z}}\right]^{-1}\left((\boldsymbol{L}^T\boldsymbol{L})^{-1}_{\mathcal{Z}\mathcal{S}}(\boldsymbol{L}^T_{\mathcal{S}}\boldsymbol{y} - \frac{\sigma^2}{\phi}\bar{\boldsymbol{v}}_{\mathcal{S}})\right).$$

Similarly

$$\begin{aligned}\bar{\boldsymbol{\beta}}_{\mathcal{S}} &= (\boldsymbol{L}^T\boldsymbol{L})^{-1}_{\mathcal{S}\mathcal{S}}\boldsymbol{L}^T_{\mathcal{S}}\boldsymbol{y} + (\boldsymbol{L}^T\boldsymbol{L})^{-1}_{\mathcal{S}\mathcal{Z}}\boldsymbol{L}^T_{\mathcal{Z}}\boldsymbol{y} - \frac{\sigma^2}{\phi}(\boldsymbol{L}^T\boldsymbol{L})^{-1}_{\mathcal{S}\mathcal{S}}\bar{\boldsymbol{v}}_{\mathcal{S}} - \frac{\sigma^2}{\phi}(\boldsymbol{L}^T\boldsymbol{L})^{-1}_{\mathcal{S}\mathcal{Z}}\bar{\boldsymbol{v}}_{\mathcal{Z}}\\ &= (\boldsymbol{L}^T\boldsymbol{L})^{-1}_{\mathcal{S}\mathcal{S}}\boldsymbol{L}^T_{\mathcal{S}}\boldsymbol{y} - \frac{\sigma^2}{\phi}(\boldsymbol{L}^T\boldsymbol{L})^{-1}_{\mathcal{S}\mathcal{S}}\bar{\boldsymbol{v}}_{\mathcal{S}} - (\boldsymbol{L}^T\boldsymbol{L})^{-1}_{\mathcal{S}\mathcal{Z}}\left[(\boldsymbol{L}^T\boldsymbol{L})^{-1}_{\mathcal{Z}\mathcal{Z}}\right]^{-1}\left((\boldsymbol{L}^T\boldsymbol{L})^{-1}_{\mathcal{Z}\mathcal{S}}(\boldsymbol{L}^T_{\mathcal{S}}\boldsymbol{y} - \frac{\sigma^2}{\phi}\bar{\boldsymbol{v}}_{\mathcal{S}})\right)\\ &= \left((\boldsymbol{L}^T\boldsymbol{L})^{-1}_{\mathcal{S}\mathcal{S}} - (\boldsymbol{L}^T\boldsymbol{L})^{-1}_{\mathcal{S}\mathcal{Z}}\left[(\boldsymbol{L}^T\boldsymbol{L})^{-1}_{\mathcal{Z}\mathcal{Z}}\right]^{-1}(\boldsymbol{L}^T\boldsymbol{L})^{-1}_{\mathcal{Z}\mathcal{S}}\right)\left(\boldsymbol{L}^T_{\mathcal{S}}\boldsymbol{y} - \frac{\sigma^2}{\phi}\bar{\boldsymbol{v}}_{\mathcal{S}}\right)\\ &= \left[(\boldsymbol{L}^T\boldsymbol{L})^{-1}_{\mathcal{S}\mathcal{S}}\right]^{-1}\left(\boldsymbol{L}^T_{\mathcal{S}}\boldsymbol{y} - \frac{\sigma^2}{\phi}\bar{\boldsymbol{v}}_{\mathcal{S}}\right)\\ &= (\boldsymbol{L}^T_{\mathcal{S}}\boldsymbol{L}_{\mathcal{S}})^{-1}\boldsymbol{L}^T_{\mathcal{S}}\boldsymbol{y} - \frac{\sigma^2}{\phi}(\boldsymbol{L}^T_{\mathcal{S}}\boldsymbol{L}_{\mathcal{S}})^{-1}\bar{\boldsymbol{v}}_{\mathcal{S}}.\end{aligned}$$

□

This theorem implies that LASSO computation can be significantly reduced if the partition into the sets $\mathcal{S}$ and $\mathcal{Z}$ are known. This is implicitly exploited in the interior point descent algorithm (Chapter 4) which constructs the non-zero set $\mathcal{S}$ incrementally. The matrix $(\boldsymbol{L}^T_{\mathcal{S}}\boldsymbol{L}_{\mathcal{S}})^{-1}$ can be constructed incrementally in this algorithm removing a substantial amount of computing. This strategy in described in more detail in Chapter 6.

5.5 Summary

In this chapter some of the important distributions associated with the simplest (pathological) LASSO model have been investigated (Section 5.2). The marginal distribution of the outcomes for the multiple effect LASSO model was also derived (Theorem 5.6). This distribution is difficult to calculate and/or maximise. Hence, alternative expressions are needed. Finally, it was shown that computation for the LASSO predictions can be simplified if the partition into zero and non-zero estimates is known prior to analysis.

Chapter 6

LASSO Computing

6.1 Introduction

The computing methods for estimating LASSO effects, described in Chapter 4, form the basis for the computing methods used throughout this thesis. However, as described in Chapter 4 these methods are not sufficiently flexible to perform some of the tasks that will be required in later chapters. In particular, the interior point descent algorithm described in Chapter 4 assumes that, conditional on the LASSO random effects, the observed outcomes are independent and identically distributed. In this chapter this requirement is relaxed to allow a known covariance structure.

The interior point descent algorithm for LASSO effect estimation is implemented using an efficient updating routine, outlined in this chapter. For a given constraint parameter, the algorithm starts with a null set of explanatory variables to which a single variable is added after each iteration. This addition is performed without the need to repeat many calculations at each iteration.

When introducing the interior point algorithm, Osborne et al. (2000) suggested that the penalised regression (4.5.1) could be estimated using the same methods for estimation of the constrained regression. They suggested numerical methods using a grid search or a Newton-Raphson algorithm. The details of a Newton-Raphson algorithm are described in this chapter.

The S-PLUS 6.1 functions to implement the interior point descent algorithm for correlated data and the penalised regression are given in Appendix D. Faster implementations are nearly always possible. For example it is expected that implementation in a lower level language such as Fortran or C++ would speed up computation significantly.

Finally, this chapter includes a discussion of the generalised cross validation (GCV) style statistic of Tibshirani (1996) and Fu (1998). The estimate of the degrees of freedom for the LASSO model given in Tibshirani (1996) is not monotonic and has discontinuities when considered as a function of t. In contrast, the estimate of Fu (1998) appears to be monotonic and may provide a better method for choosing the constraint or penalty parameter.

6.2 Interior Point Descent Algorithm for Correlated Data

Recall the LASSO random linear model

$$\boldsymbol{y} = \boldsymbol{L}\boldsymbol{\beta} + \boldsymbol{e}$$

where $\boldsymbol{y}$ is the $n \times 1$ vector of observed outcomes, $\boldsymbol{\beta}$ is the $q \times 1$ vector of LASSO effects, $\boldsymbol{L}$ is the $n \times q$ design matrix and $\boldsymbol{e}$ is the $n \times 1$ vector of residuals. Up to this point in the thesis, the elements of $\boldsymbol{e}$ are assumed to be independently and identically distributed. This assumption is now relaxed by allowing $\operatorname{var}(\boldsymbol{e}) = \sigma^2\boldsymbol{R}$ where $\boldsymbol{R}$ is an arbitrary correlation structure. The LASSO optimisation problem for independent and identically distributed residuals in Chapter 4 can be extended to allow for this known correlation structure.

$$\begin{aligned} &\underset{\boldsymbol{\beta}\in\mathbb{R}^q}{\text{minimise:}} && S(\boldsymbol{\beta}) = \frac{1}{2}(\boldsymbol{y} - \boldsymbol{L}\boldsymbol{\beta})^T\boldsymbol{R}^{-1}(\boldsymbol{y} - \boldsymbol{L}\boldsymbol{\beta}) \\ &\text{subject to:} && \|\boldsymbol{\beta}\|_1 \leq t \end{aligned}$$

where $0 \leq t < \|\hat{\boldsymbol{\beta}}\|_1$ and $\hat{\boldsymbol{\beta}} = (\boldsymbol{L}^T\boldsymbol{R}^{-1}\boldsymbol{L})^{-1}\boldsymbol{L}^T\boldsymbol{R}^{-1}\boldsymbol{y}$ is the unconstrained estimate.

The incorporation of the correlation matrix $\boldsymbol{R}$ into the optimisation problem causes change in the equations used in the interior point algorithm but not in the iterative process. The derivation of the equations needed for the algorithm is now given. It follows closely the method described in Osborne et al. (2000) for independent and identically distributed residuals, given in Chapter 4.

6.2.1 *First Order Conditions*

The Lagrangian of the constraint problem is the penalised regression

$$\mathscr{L}(\boldsymbol{\beta}, \lambda) = \frac{1}{2}(\boldsymbol{y} - \boldsymbol{L}\boldsymbol{\beta})^T\boldsymbol{R}^{-1}(\boldsymbol{y} - \boldsymbol{L}\boldsymbol{\beta}) + \lambda\|\boldsymbol{\beta}\|_1 - \lambda t \tag{6.2.1}$$

where λ is the Lagrange multiplier corresponding to constraint t. The sub-differential of the Lagrangian for the correlated constraint problem is

$$\partial_{\boldsymbol{\beta}}(\boldsymbol{\beta}, \lambda) = -\boldsymbol{L}^T\boldsymbol{R}^{-1}(\boldsymbol{y} - \boldsymbol{L}\boldsymbol{\beta}) + \lambda\boldsymbol{v}$$

where $\boldsymbol{v}$ is a vector whose elements, v_i, are elements of $[-1, 1]$ such that $\boldsymbol{v}^T\boldsymbol{\beta} = \|\boldsymbol{\beta}\|_1$. At a solution $\bar{\boldsymbol{\beta}}$ to the Lagrangian $\boldsymbol{0} \in \partial_{\boldsymbol{\beta}}(\boldsymbol{\beta}, \lambda)$ and

$$\boldsymbol{0} = -\boldsymbol{L}^T\boldsymbol{R}^{-1}(\boldsymbol{y} - \boldsymbol{L}\bar{\boldsymbol{\beta}}) + \lambda\bar{\boldsymbol{v}} \tag{6.2.2}$$

so that

$$\bar{\boldsymbol{\beta}} = (\boldsymbol{L}^T\boldsymbol{R}^{-1}\boldsymbol{L})^{-}\boldsymbol{L}^T\boldsymbol{R}^{-1}\boldsymbol{y} - \lambda(\boldsymbol{L}^T\boldsymbol{R}^{-1}\boldsymbol{L})^{-}\bar{\boldsymbol{v}}$$

where $\bar{\boldsymbol{v}}$ is a realisation of $\boldsymbol{v}$ such that $\bar{\boldsymbol{v}}^T\bar{\boldsymbol{\beta}} = \|\bar{\boldsymbol{\beta}}\|_1$. Taking the L_∞-norm of the estimating equation in (6.2.2) yields the expression for λ

$$\lambda\bar{\boldsymbol{v}} = \boldsymbol{L}^T\boldsymbol{R}^{-1}(\boldsymbol{y} - \boldsymbol{L}\bar{\boldsymbol{\beta}})$$

so that

$$\lambda = \|\boldsymbol{L}^T\boldsymbol{R}^{-1}(\boldsymbol{y} - \boldsymbol{L}\bar{\boldsymbol{\beta}})\|_\infty.$$

These expressions in turn allow the form of $\bar{\boldsymbol{v}}$ to be specified.

$$\bar{\boldsymbol{v}} = \frac{\boldsymbol{L}^T\boldsymbol{R}^{-1}(\boldsymbol{y} - \boldsymbol{L}^T\bar{\boldsymbol{\beta}})}{\|\boldsymbol{L}^T\boldsymbol{R}^{-1}(\boldsymbol{y} - \boldsymbol{L}\bar{\boldsymbol{\beta}})\|_\infty}.$$

This equation gives an optimality check for the algorithm to be discussed presently. If $\bar{\boldsymbol{v}}$ satisfies $\|\bar{\boldsymbol{v}}\|_\infty = 1$ then the solution has been found (Osborne et al., 2000).

6.2.2 *Linearisation and Optimal Descent Directions*

The optimal descent directions given for the independent and identically distributed residual case in Chapter 4 are easily extended to the correlated residual case. Paralleling the independent and identically distributed residual case, the descent directions are those that satisfy the Karush-Kuhn-Tucker conditions in Result C.1. Let the descent direction for the current estimate be $\boldsymbol{h}^{(k)}$ and the descent direction for the Lagrange multiplier be $\nu^{(k)}$. The descent directions for $\boldsymbol{\beta}_{\mathcal{S}}$ and λ are given as the solutions to

$$\begin{pmatrix} \boldsymbol{L}_{\mathcal{S}}^T\boldsymbol{R}^{-1}\boldsymbol{L}_{\mathcal{S}} & \bar{\boldsymbol{v}}_{\mathcal{S}}^{(k)} \\ \bar{\boldsymbol{v}}_{\mathcal{S}}^{(k)^T} & \mathbf{0} \end{pmatrix} \begin{pmatrix} \boldsymbol{h}_{\mathcal{S}}^{(k)} \\ \nu^{(k)} \end{pmatrix} = \begin{pmatrix} \boldsymbol{L}_{\mathcal{S}}^T\boldsymbol{R}^{-1}(\boldsymbol{y} - \boldsymbol{L}_{\mathcal{S}}\bar{\boldsymbol{\beta}}_{\mathcal{S}}^{(k)}) \\ t - \bar{\boldsymbol{v}}_{\mathcal{S}}^{(k)^T}\bar{\boldsymbol{\beta}}_{\mathcal{S}}^{(k)} \end{pmatrix}.$$

The superscripts have been added to show that the solution to these equations is only local and iteration is needed to find the global solution (see Result C.1).

6.2.3 *Interior Point Descent Algorithm*

Mirroring the interior point descent algorithm for the independent and identically distributed residuals (Chapter 4) the LASSO model effects $\boldsymbol{\beta}$ are divided into those that are estimated to be non-zero ($\boldsymbol{\beta}_{\mathcal{S}}$) and those that are estimated to be zero ($\boldsymbol{\beta}_{\mathcal{Z}}$). Let $\boldsymbol{\theta}^{(k)^T} = (\boldsymbol{\theta}_{\mathcal{S}}^{(k)^T}, \mathbf{0}^T) = (\text{sign}(\bar{\boldsymbol{\beta}}_{\mathcal{S}})^{(k)^T}, \mathbf{0}^T)$ be the current vector of the signs of $\bar{\boldsymbol{\beta}}$. The descent algorithm updates the current non-zero LASSO estimates $\bar{\boldsymbol{\beta}}_{\mathcal{S}}^{(k)}$ by

$$\begin{aligned} \bar{\boldsymbol{\beta}}_{\mathcal{S}}^{(k+1)} &= \bar{\boldsymbol{\beta}}_{\mathcal{S}}^{(k)} + \boldsymbol{h}_{\mathcal{S}}^{(k)} \qquad \text{where} \\ \boldsymbol{h}_{\mathcal{S}}^{(k)} &= (\boldsymbol{L}_{\mathcal{S}}^T\boldsymbol{R}^{-1}\boldsymbol{L}_{\mathcal{S}})^{-1}(\boldsymbol{L}_{\mathcal{S}}^T\boldsymbol{R}^{-1}(\boldsymbol{y} - \boldsymbol{L}_{\mathcal{S}}\bar{\boldsymbol{\beta}}_{\mathcal{S}}^{(k)}) - \nu^{(k)}\boldsymbol{\theta}_{\mathcal{S}}^{(k)}) \qquad \text{and} \\ \nu^{(k)} &= \max\left(0, \frac{\boldsymbol{\theta}_{\mathcal{S}}^{(k)^T}(\boldsymbol{L}_{\mathcal{S}}^T\boldsymbol{R}^{-1}\boldsymbol{L}_{\mathcal{S}})^{-1}\boldsymbol{L}_{\mathcal{S}}^T\boldsymbol{R}^{-1}\boldsymbol{y} - t}{\boldsymbol{\theta}_{\mathcal{S}}^{(k)^T}(\boldsymbol{L}_{\mathcal{S}}^T\boldsymbol{R}^{-1}\boldsymbol{L}_{\mathcal{S}})^{-1}\boldsymbol{\theta}_{\mathcal{S}}^{(k)}}\right). \end{aligned} \tag{6.2.3}$$

These generalised forms of the updating equations for the interior point descent algorithm are now incorporated into the iterative procedure in Chapter 4.

6.2.4 *Variable Updating*

The interior point descent algorithm adds a new variable to the active set $\mathcal{S}$ after each iteration. Hence, the algorithm behaves similarly to a forward moving algorithm. Some of the computational advantages used in single variable addition (e.g. Seber, 1977) can be employed in the LASSO estimation process. A QR factorisation method is suggested in Osborne et al. (2000) but no details are given. Here the details of an alternative, simple and perhaps numerically naïve method are given.

After the k^{th} iteration, there will be at most k variables in the active set $\mathcal{S}^{(k)}$ and it is guaranteed that there are at most n non-zero LASSO estimates, where n is the number of observations (Osborne et al., 2000, Theorem 6). Without loss of generality assume that $|\mathcal{S}^{(k)}| < n$, otherwise the solution would have been found. To obtain the LASSO estimate for the k^{th} iteration the matrix $(\boldsymbol{L}^T_{\mathcal{S}^{(k)}}\boldsymbol{R}^{-1}\boldsymbol{L}_{\mathcal{S}^{(k)}})^{-1}$ is calculated. Assume the $(k+1)^{th}$ iteration identifies variable $\boldsymbol{l}_j$ as being the next variable to enter the active set. To calculate the LASSO effects, the matrix $(\boldsymbol{L}^T_{\mathcal{S}^{(k+1)}}\boldsymbol{R}^{-1}\boldsymbol{L}_{\mathcal{S}^{(k+1)}})^{-1}$ is needed. This can be expressed as a partitioned matrix and inverted based on the previous iteration's solution by using Result B.4.

$$\begin{aligned}(\boldsymbol{L}^T_{\mathcal{S}^{(k+1)}}\boldsymbol{R}^{-1}\boldsymbol{L}_{\mathcal{S}^{(k+1)}})^{-1} &= \begin{pmatrix} \boldsymbol{L}^T_{\mathcal{S}^{(k)}}\boldsymbol{R}^{-1}\boldsymbol{L}_{\mathcal{S}^{(k)}} & \boldsymbol{L}^T_{\mathcal{S}^{(k)}}\boldsymbol{R}^{-1}\boldsymbol{l}_j \\ \boldsymbol{l}_j^T\boldsymbol{R}^{-1}\boldsymbol{L}_{\mathcal{S}^{(k)}} & \boldsymbol{l}_j^T\boldsymbol{R}^{-1}\boldsymbol{l}_j \end{pmatrix} \\ &= \begin{pmatrix} (\boldsymbol{L}^T_{\mathcal{S}^{(k)}}\boldsymbol{R}^{-1}\boldsymbol{L}_{\mathcal{S}^{(k)}})^{-1} + \boldsymbol{B}\boldsymbol{T}\boldsymbol{B}^T & -\boldsymbol{T}\boldsymbol{B}^T \\ -\boldsymbol{B}\boldsymbol{T} & \boldsymbol{T} \end{pmatrix}^{-1}\end{aligned}$$

where

$$\begin{aligned}\boldsymbol{T} &= \left(\boldsymbol{l}_j^T\boldsymbol{R}^{-1}\boldsymbol{l} - \boldsymbol{l}_j^T\boldsymbol{R}^{-1}\boldsymbol{L}_{\mathcal{S}^{(k)}}(\boldsymbol{L}^T_{\mathcal{S}^{(k)}}\boldsymbol{R}^{-1}\boldsymbol{L}_{\mathcal{S}^{(k)}})^{-1}\boldsymbol{L}_{\mathcal{S}^{(k)}}\boldsymbol{R}^{-1}\boldsymbol{l}\right)^{-1} \quad \text{and} \\ \boldsymbol{B} &= (\boldsymbol{L}^T_{\mathcal{S}^{(k)}}\boldsymbol{R}^{-1}\boldsymbol{L}_{\mathcal{S}^{(k)}})^{-1}\boldsymbol{L}_{\mathcal{S}^{(k)}}\boldsymbol{R}^{-1}\boldsymbol{l}.\end{aligned}$$

The matrices $\boldsymbol{T}$ and $\boldsymbol{B}$ are calculated with minimal computational cost since $(\boldsymbol{L}^T_{\mathcal{S}^{(k)}}\boldsymbol{R}^{-1}\boldsymbol{L}_{\mathcal{S}^{(k)}})^{-1}$ is already known from the previous iteration.

During the sign change step of the algorithm a variable is removed from the active set. In a typical run of the interior point algorithm the number of up-dating steps far outweighs the number of down-dating steps and hence the method used to down-date is not as important as the method to up-date. There is no analogous routine to that described for up-dating for down-dating the matrix $(\boldsymbol{L}^T_{\mathcal{S}^{(k)}}\boldsymbol{R}^{-1}\boldsymbol{L}_{\mathcal{S}^{(k)}})^{-1}$. A decomposition of this matrix, such as the QR decomposition, could be utilised to increase efficiency but has not been implemented. Whenever a down-dating step is needed, the complete matrix is inverted. This is a computationally expensive but correct approach.

6.3 Estimating Penalised Regression

The interior point descent algorithm can be used to find the solution to the penalised regression (6.2.1). As discussed in Section 6.2.1 the penalised regression equation is the Lagrangian for the constraint problem and the penalty parameter λ is the Lagrange multiplier. From a statistical viewpoint, the task of estimating the penalised regression is equivalent to finding the t for which the Lagrange multiplier is λ.

There is a penalty λ_{max} such that all $\lambda \geq \lambda_{max}$ induces all LASSO estimates to be identically zero, that is they have constraint $t = 0$. Without loss of generality, the set of possible λ is reduced to $\mathscr{P} = \{\lambda | 0 \leq \lambda \leq \lambda_{max}\}$. Since any $\lambda \in \mathscr{P}$ is a Lagrange multiplier for a particular $t \in \mathscr{T} = \{t | 0 \leq t \leq \|\hat{\boldsymbol{\beta}}\|_1\}$ then there is a relationship between $t \in \mathscr{T}$ and $\lambda \in \mathscr{P}$. Further, for each $t \in \mathscr{T}$ there must be one and only one corresponding $\lambda \in \mathscr{P}$, that is the relationship between $t \in \mathscr{T}$ and $\lambda \in \mathscr{P}$ is a function $h(t) = \lambda$. The function must be surjective. If it were not, then $\mathscr{T}$ would not completely define the set of possible LASSO solutions. The function $h(t)$ is monotonic decreasing as increasing the constraint t implies that there is less penalty λ on the unconstrained estimating equation. Further, it is strictly monotonic decreasing - if it were not then the LASSO solutions, defined by $\lambda \in \mathscr{P}$, would not be unique.

The function $h(t)$ is easily evaluated by finding the LASSO solution and then calculating the Lagrange multiplier for that solution. An example of the function $h(t) = \lambda$ is given in Figure 6.1 for the prostate data from Stamey et al. (1989). Note that it has all the properties described.

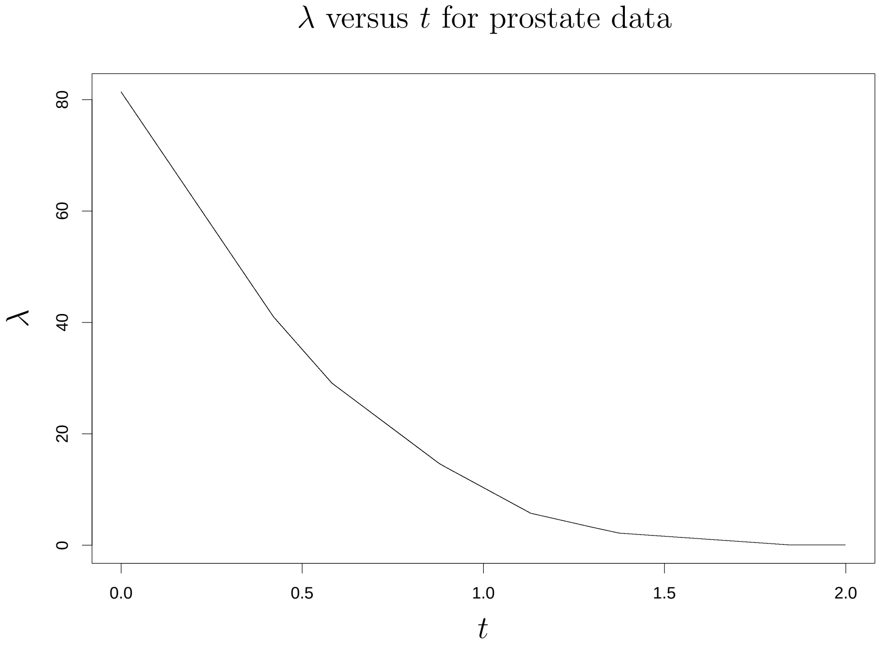

Figure 6.1: Penalty parameter λ versus constraint parameter t for LASSO models fitted to the prostate data.

To solve the penalised regression using the interior point descent algorithm for a given penalty, λ_g say, the constraint $t \in \mathscr{T}$ such that $h(t) = \lambda_g$ is required. This is achieved by the iterative Newton-Raphson procedure. Each iteration is updated via

$$t^{(i+1)} = t^{(i)} - \frac{h\left(t^{(i)}\right) - \lambda_g}{h'\left(t^{(i)}\right)}$$

where $t^{(i)}$ is the previous iteration and $h'(t)$ is the derivative of $h(t)$. Iteration is carried out until convergence with some predefined tolerance is achieved. The functional form of $h(t)$ is unknown but its value at any point can be calculated - it is the value of the Lagrange multiplier at the LASSO solution for constraint t. The value of $h'(t)$ is also unknown and must be approximated. The forward difference approximation to the differential

$$h'(t) \approx \frac{h(t) - h(t + \Delta t)}{\Delta t}$$

is used, where Δt is a small positive number. If the limit as $\Delta t \rightarrow 0$ of this approximation is taken, then it is the differential (tangent) of the function h at t. This iterative method is called a *numerical* Newton-Raphson algorithm as the derivative (or differential) is approximated numerically.

Using the interior point descent algorithm, models corresponding to a small t are quicker to fit than those with a large t. This makes $t = 0$ an obvious starting point for the numerical Newton-Raphson search.

Care must be taken with all numerical methods when choosing the convergence tolerance and arbitrary constants used in the computation. In practice, for all data sets encountered so far, a convergence tolerance of 0.001 and $\Delta t = 0.000001$ has been used and seems to perform acceptably. If the convergence tolerance is decreased, then the value of Δt should also be decreased and under no circumstances should Δt be larger than the convergence tolerance. If it was larger, then it is possible to obtain $t^{(i)} < t_g < t^{(i)} + \Delta t$, which could cause convergence failure.

6.3.1 *Computation Costs for Penalised Regression*

The computational cost of calculating the approximate derivative is minimal. The LASSO solution used in calculating $h\left(t^{(i)}\right)$ forms an interior point for $h\left(t^{(i)} + \Delta t\right)$ which is extremely close to the new solution and is almost guaranteed to be reached with one step of the interior point descent algorithm.

Calculating $h\left(t^{(i)}\right)$ requires considerably more computation, especially if the set of active explanatory variables in the descent algorithm is started at the null set for every iteration. The computation cost can be significantly reduced by identifying better starting values for the descent algorithm. A method for doing this is now described.

The Newton-Raphson search is started at $t = 0$, the first movement of the search increases t, as will any movement from a point $t^{(i)} < t_g$ where $h(t_g) = \lambda_g$. For these steps the previous iteration's solution to the interior point descent algorithm can be used as starting values. The movement from a point where $t^{(i)} > t_g$ will decrease $t^{(i)}$. This means that the previous iteration's solution cannot be used as the starting point as it is not an interior point. Instead the solution to the k^{th} iteration is used where k is the largest value such that $t^{(k)} \leq t_g$ or equivalently $h\left(t^{(k)}\right) \geq h(t_g) = \lambda_g$. This point is guaranteed to be an interior point for the current iteration. It turns out that the choice of k can always be taken to be the iteration immediately prior to the previous iteration. The proof is given in the following Theorem (to demonstrate that each step is closer to the solution) and Corollary (showing the point just raised).

Theorem 6.1. *Let $h(t)$ be a non-negative strictly monotonic decreasing function of t. The solution to $h(t) = \lambda_g$ is required. Let t_g be this solution. If a Newton-Raphson iterative search is employed by defining $t^{(i+1)} = t^{(i)} - \frac{h\left(t^{(i)}\right)-\lambda_g}{h'\left(t^{(i)}\right)}$ where $h'(t)$ is the differential of h at t then*

$$|t^{(i+1)} - t_g| \leq |t^{(i)} - t_g|.$$

That is the updated $t^{(i+1)}$ is closer to the solution than the previous $t^{(i)}$.

Proof. Note that since $h(t)$ is strictly monotonic decreasing then $h'(t) < 0$ for all t. Now

$$\begin{aligned} \left|t^{(i+1)} - t_g\right| &= \left|t^{(i+1)} - t^{(i)} + t^{(i)} - t_g\right| \\ &= \left|-\frac{h\left(t^{(i)}\right) - \lambda_g}{h'\left(t^{(i)}\right)} + t^{(i)} - t_g\right|. \end{aligned}$$

If $t^{(i)} - t_g > 0$ then $-\frac{h\left(t^{(i)}\right)-\lambda_g}{h'\left(t^{(i)}\right)} < 0$ and

$$\left|t^{(i+1)} - t_g\right| = \left|-\frac{h\left(t^{(i)}\right) - \lambda_g}{h'\left(t^{(i)}\right)} + t^{(i)} - t_g\right| < \left|t^{(i)} - t_g\right|.$$

If $t^{(i)} - t_g < 0$ then $-\frac{h\left(t^{(i)}\right)-\lambda_g}{h'\left(t^{(i)}\right)} > 0$ and

$$\left|t^{(i+1)} - t_g\right| = \left|-\frac{h\left(t^{(i)}\right) - \lambda_g}{h'\left(t^{(i)}\right)} + t^{(i)} - t_g\right| < \left|t^{(i)} - t_g\right|.$$

In summary

$$\left|t^{(i+1)} - t_g\right| \leq \left|t^{(i)} - t_g\right|.$$

□

Theorem 6.1 guarantees convergence if the differential is known, as each updated value is consecutively closer to the solution. When the differential is replaced by an approximation, convergence is not guaranteed. However, if the approximation is good, then convergence is almost always expected.

Corollary 6.1. *Let $t^{(k)} < t_g$ be such that $t^{(k+1)} > t_g$ then $t^{(k)} \leq t^{(i)}$ for all $i > k$.*

Proof. If $t^{(i)} > t_g$ then the result is immediate. If $t^{(i)} < t_g$ then from Theorem 6.1

$$t^{(k)} - t_g < t^{(i)} - t_g \qquad \text{for all } i > k$$

giving

$$t^{(k)} < t^{(i)}$$

as $t^{(i)} = t^{(k+j)}$, $j = 1, 2, \ldots$ □

Using this simple scheme for generating starting points for the interior point descent algorithm reduces the computation cost of fitting penalised regressions significantly. In fact, if there are not too many backward steps, those with $t^{(i+1)} < t^{(i)}$, the cost of computation should not be impedingly more than the constraint problem. For the prostate data of Stamey et al. (1989) analysed in Tibshirani (1996) the maximum difference, over t, between the two estimation methods is under 0.4 seconds, or approximately 3 times slower. The time taken to compute the solution for both the estimation methods is plotted against the constraint in Figure 6.2.

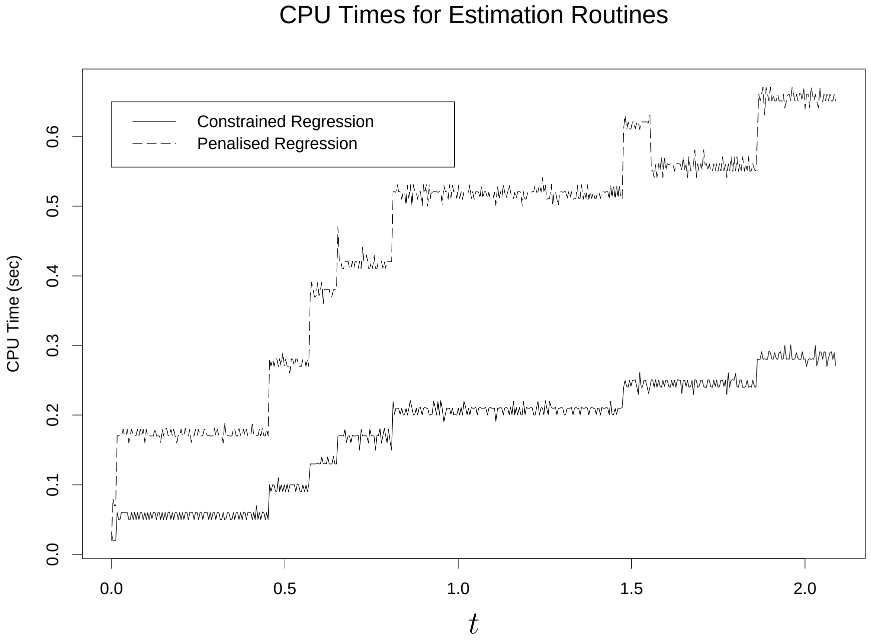

Figure 6.2: Computation times for constrained regression and penalised regression versus constraint value t. The data are the prostate data and t varies from 0 to $\|\hat{\boldsymbol{\beta}}\|_1$. Analyses performed on a laptop Pentium 4 P.C. with 512 MB RAM and a clock-speed of 2.0 GHz.

The maximum time taken for fitting models with a large number of explanatory variables is greater than that for a model with a smaller number of explanatory variables. This applies to both the constrained and penalised regression. The difference in computing time for the constrained and penalised regressions is expected to grow for large models as the number of descent steps is greater.

6.4 Discussion of Degrees of Freedom for GCV

A generalized cross validation (GCV) style statistic was proposed by Tibshirani (1996) as a method of choosing the constraint parameter t. A discussion on methods to calculate the statistic is now given. The appropriateness of using the GCV to choose the constraint is not considered. Much of the content of the forthcoming chapters presents a different method for choosing the constraint, one that the author feels is more natural as it is based on a statistical model. The GCV statistic is based on the ridge regression approximation (Tibshirani, 1996)

$$\bar{\boldsymbol{\beta}} = (\boldsymbol{L}^T\boldsymbol{L} + \lambda\boldsymbol{W}^-)^{-1}\boldsymbol{L}^T\boldsymbol{y} \tag{6.4.1}$$

where λ is the Lagrange multiplier or penalty parameter, $\boldsymbol{W}$ is a diagonal matrix with $|\bar{\boldsymbol{\beta}}|$ on its diagonal and $\boldsymbol{W}^-$ is a generalised inverse; here it is the Moore-Penrose inverse. Using the ridge regression approximation the GCV style statistic given by Tibshirani (1996) is

$$\mathrm{GCV}(t) = \frac{RSS(t)}{n\left(1 - \frac{q(t)}{n}\right)^2}$$

where

$$\begin{aligned} q(t) &= \mathrm{tr}\left(\boldsymbol{L}\left(\boldsymbol{L}^T\boldsymbol{L} + \lambda\boldsymbol{W}^-\right)^{-1}\boldsymbol{L}^T\right) \\ &= \mathrm{tr}\left(\boldsymbol{L}^T\boldsymbol{L}\left(\boldsymbol{L}^T\boldsymbol{L} + \lambda\boldsymbol{W}^-\right)^{-1}\right) \end{aligned}$$

$RSS(t)$ is the residual sums of squares for the constraint t and $q(t)$ is the number of *effective* parameters or the *effective* degrees of freedom in the fitted model.

Theorem 6.2. *Let $\boldsymbol{L}^T\boldsymbol{L}$ be positive definite. Then the degrees of freedom estimate $q(t)$ from Tibshirani (1996) can be expressed as*

$$q(t) = \sum_{i=1}^{q} \frac{1}{1+e_i} = q_{\mathcal{Z}} + \sum_{i=1}^{q_{\mathcal{S}}} \frac{1}{1+e_i^*}$$

where q is the number of columns in the design matrix $\boldsymbol{L}$, $q_{\mathcal{S}}$ is the number of non-zero LASSO predictions, $q_{\mathcal{Z}}$ is the number of zero LASSO predictions, e_i is the i^{th} eigenvalue of $\lambda\boldsymbol{W}^-(\boldsymbol{L}^T\boldsymbol{L})^{-1}$ and e_i^* is the i^{th} non-zero eigenvalue of $\lambda\boldsymbol{W}^-(\boldsymbol{L}^T\boldsymbol{L})^{-1}$.

Proof.

$$\begin{aligned} q(t) &= \mathrm{tr}\left(\boldsymbol{L}^T\boldsymbol{L}(\boldsymbol{L}^T\boldsymbol{L} + \lambda\boldsymbol{W}^-)^{-1}\right) \\ &= \mathrm{tr}\left((\boldsymbol{I} + \lambda\boldsymbol{W}^-(\boldsymbol{L}^T\boldsymbol{L})^{-1})^{-1}\right) \\ &= \sum_{i=1}^{q} \frac{1}{1+e_i}. \end{aligned}$$

There will be $q_{\mathcal{Z}}$ zero eigenvalues of $\boldsymbol{W}$ and hence $\lambda \boldsymbol{W}^{-}(\boldsymbol{L}^T\boldsymbol{L})^{-1}$. So

$$q(t) = q_{\mathcal{Z}} + \sum_{i=1}^{q_{\mathcal{S}}} \frac{1}{1+e_i^*}.$$

□

Intuitively, it is expected that if there is a severe constraint placed on the estimates then $q(t) = 0$. Further, $q(t)$ should be a monotonic increasing function of t. This is not the case; when $t = 0$ the matrix $\boldsymbol{W}^{-}$ is identically zero, as are all the eigenvalues of $\lambda \boldsymbol{W}^{-}(\boldsymbol{L}^T\boldsymbol{L})^{-1}$ and $q(t) = q$. This is the maximum value of $q(t)$ and hence it cannot be a monotonically increasing function.

An alternative definition of $q(t)$ was given in Fu (1998). In this definition the number of zero estimates is subtracted from the previous definition. That is

$$q_f(t) = \operatorname{tr}\left(\boldsymbol{L}^T\boldsymbol{L}\left(\boldsymbol{L}^T\boldsymbol{L} + \lambda \boldsymbol{W}^{-}\right)^{-1}\right) - q_{\mathcal{Z}}$$

where $q_{\mathcal{Z}}$ is the number of zero estimates. Fu (1998) states that this adjustment to the number of effective parameters is performed to compensate for the loss of the diagonals of $\boldsymbol{W}$ due to zero estimates. This is evident from Theorem 6.2. The function $q_f(t)$ can be expressed as

$$q_f(t) = \sum_{i=1}^{q_{\mathcal{S}}} \frac{1}{1+e_i^*}.$$

The effective degrees of freedom estimate $q_f(t)$ has attributes $q_f(0) = 0$ and $q_f(\|\hat{\boldsymbol{\beta}}\|_1) = q$ which makes it more appealing than $q(t)$. It is unclear if it is monotonic increasing for the rest of the domain, although empirical evidence suggests that it is. Plots of $q(t)$ and $q_f(t)$ for the prostate data from Stamey et al. (1989) are given in Figure 6.3.

The function $q(t)$ has discontinuities at the points where the number of non-zero estimates change. This arises as the rank of the matrix $\boldsymbol{W}$ also changes at these points. The function $q_f(t)$ also has small discontinuities at least at some of these points. They are visible in Figure 6.3 at the model sizes 5 and 7. The size of the jumps at the discontinuities in $q_f(t)$ are reduced drastically and are reversed in sign when compared to those in $q(t)$.

6.4.1 *Impact of Estimated Degrees of Freedom on GCV Statistic*

The GCV statistic depends heavily on the value of the effective degrees of freedom estimate. For GCV to be a reasonable method for choosing the constraint parameters the effective degrees of freedom estimator should be an increasing function of the constraint parameter. The effective degrees of freedom estimate in Tibshirani (1996) does not have this property and hence may produce GCV statistics that are unpredictable. The GCV statistic based on $q_f(t)$ from Fu (1998) has this property and may perform better. The simulation studies in Tibshirani (1996) suggest that the GCV method for choosing the constraint provides, over a number of different model setups, good results for prediction. However, the preceding discussion showed that GCV based on $q_f(t)$ is likely to perform better.

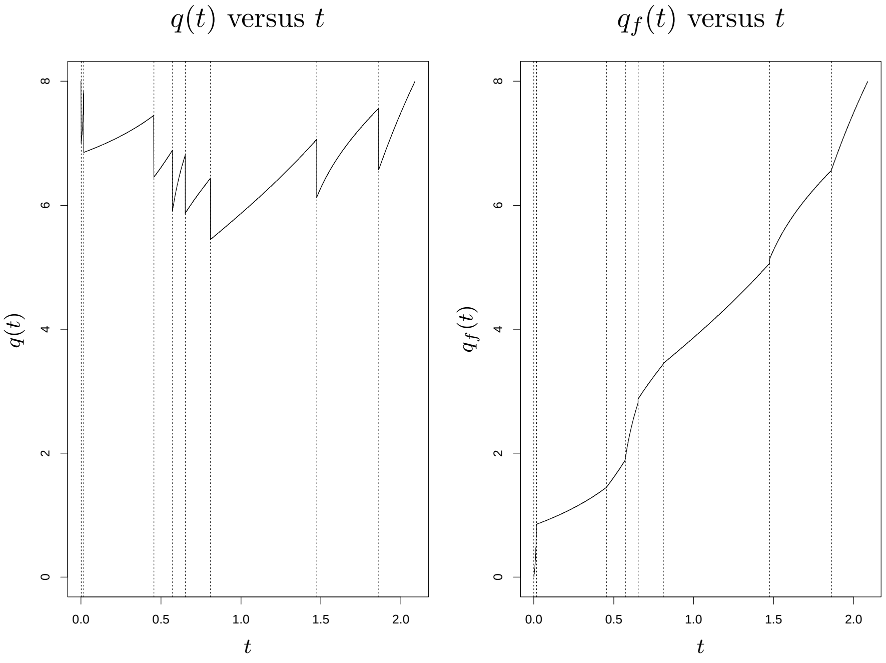

Figure 6.3: Effective degrees of freedom estimates. Dotted vertical lines indicate where the number of non-zero estimates change.

6.5 Summary

In this chapter the base methods for computing used in this thesis were outlined. In particular, the methodology was presented for estimating the constrained and penalised regression when the outcomes are dependent. This methodology is based on the interior point descent algorithm (Chapter 4). This chapter also presented a discussion on the GCV method to estimate the constraint or penalty parameter. It was shown that the GCV statistic presented by Tibshirani (1996) is unlikely to produce sensible results.

Chapter 7

The LASSO as a Random Effects Model

7.1 Introduction

In previous chapters the LASSO model has been viewed as the random model

$$\boldsymbol{y} = \boldsymbol{L}\boldsymbol{\beta} + \boldsymbol{e} \tag{7.1.1}$$

where the effects $\boldsymbol{\beta}$ were assumed to be double exponentially distributed variables with variance $2\phi^2$ and the residuals $\boldsymbol{e}$ were assumed to be independently and identically distributed normal variables with variance σ^2. No assumption about the number of LASSO effects is made. It is possible that the number of LASSO effects (q) is greater than the number of observations (n).

Conventional statistical analysis using random effects models starts by finding the marginal distribution of the outcomes. Likelihood estimation of σ^2 and ϕ can then be carried out based on this marginal distribution. In Chapter 5, exact marginal distributions were developed for a single β (Theorem 5.1) and for a vector $\boldsymbol{\beta}$ (Theorem 5.6). However, the exact distributions are computationally expensive to calculate and relatively inaccurate, especially for high dimensional problems. For this reason, an alternative expression for the likelihood is needed.

In this chapter an approximate analytical marginal likelihood is presented. The approximation is based on Laplace's method and requires a Taylor series expansion of the exponent of the joint distribution. For the random effects model (7.1.1) the exponent is not differentiable so the sub-differential (Definition C.3) and its derivative are used in place of the first and second derivatives. The dispersion parameters can be estimated from this likelihood, but such estimates exhibit bias. The bias is alleviated by adjusting the score equations by subtracting from the observed scores estimates of the scores' expected values.

After estimation of the dispersion parameters, an estimated predictive (posterior in Bayesian terminology) distribution is specified, which is used for effect prediction. Typically, the analyst will want to assess the relative importance of the LASSO explanatory variables through the estimate of the effects. The distribution of the LASSO estimates is typically unknown, except for asymptotic results (Knight & Fu, 2000). In this chapter this is overcome by using a simulation method. Numerous data sets are simulated from a model where all the LASSO effects are assigned to zero and effects are predicted for each data set using plug-in estimates of the dispersion parameters. These predictions form the empirical null distribution, with which the observed LASSO estimates can be compared.

The unbiased estimation of the dispersion parameters is shown, by simulation, to provide a natural and very competitive method for choosing the amount of shrinkage. As an example, the method is applied to the prostate data set used in Tibshirani (1996).

All the methods of analysis presented in this chapter can be implemented using the S-PLUS code in Appendix D.

7.2 Approximate Likelihood

An approximate likelihood is generated by using a Laplace approximation (see Chapter 3) of the marginal distribution of the observed data. A standard Laplace approximation requires a Taylor series expansion around the exponent of the joint distribution of $\boldsymbol{y}$ and $\boldsymbol{\beta}$. This is not available for the LASSO model as the exponent of the joint distribution is non-differentiable. Details of the method used to obtain the approximation are given in the proof of the following theorem.

Theorem 7.1. *An approximate marginal likelihood for the linear model* (7.1.1) *is*

$$\ell(\sigma^2,\phi) = -\frac{(n-\kappa)}{2}\log\sigma^2 - q\log\phi - \frac{1}{2\sigma^2}(\boldsymbol{y}^T\boldsymbol{y} - \bar{\boldsymbol{\beta}}\boldsymbol{L}^T\boldsymbol{L}\bar{\boldsymbol{\beta}})$$

where $\kappa = \min(q,n)$.

Proof. The exponent of the integrand is

$$h(\boldsymbol{\beta},\sigma^2\phi) = -\frac{1}{2\sigma^2}(\boldsymbol{y}-\boldsymbol{L\beta})^T(\boldsymbol{y}-\boldsymbol{L\beta}) - \frac{1}{\phi}\|\boldsymbol{\beta}\|_1$$

which has sub-differential

$$\partial_{\boldsymbol{\beta}} h(\boldsymbol{\beta},\sigma^2,\phi) = \frac{1}{\sigma^2}\boldsymbol{L}^T(\boldsymbol{y}-\boldsymbol{L\beta}) - \frac{1}{\phi}\boldsymbol{v}.$$

Notice that $\boldsymbol{v} = \text{sign}(\boldsymbol{\beta})$ since the β_i are continuous random variables and $\text{P}(\beta_i = 0) = 0$; within any neighbourhood of a particular $\boldsymbol{\beta}$, $\partial\boldsymbol{v}/\partial\boldsymbol{\beta} = \boldsymbol{0}$. Hence, the sub-differential's derivative is

$$\frac{\partial(\partial_{\boldsymbol{\beta}} h(\boldsymbol{\beta},\sigma^2,\phi))}{\partial\boldsymbol{\beta}} = -\frac{1}{\sigma^2}\boldsymbol{L}^T\boldsymbol{L}.$$

The marginal distribution is

$$\begin{aligned}
f(y) &= \int_{\mathbb{R}^q} f(\boldsymbol{y}|\boldsymbol{\beta})f(\boldsymbol{\beta})\partial\boldsymbol{\beta} \\
&= \frac{1}{(2\pi\sigma^2)^{n/2}(2\phi)^q}\int_{\mathbb{R}^q} \exp\left(h(\boldsymbol{\beta},\sigma^2,\phi)\right)\partial\boldsymbol{\beta} \\
&\approx \frac{1}{(2\pi\sigma^2)^{n/2}(2\phi)^q}\exp\left(h(\bar{\boldsymbol{\beta}},\sigma^2,\phi)\right)\int_{\mathbb{R}^q} \exp\left(-\frac{1}{2\sigma^2}(\boldsymbol{\beta}-\bar{\boldsymbol{\beta}})^T\boldsymbol{L}^T\boldsymbol{L}(\boldsymbol{\beta}-\bar{\boldsymbol{\beta}})\right)\partial\boldsymbol{\beta}.
\end{aligned}$$

If $q < n$, then the integrand specifies the kernel of a multivariate normal distribution. If $q > n$, then the integrand is the kernel of a singular multivariate normal distribution with singular variance matrix $\boldsymbol{K}^- = (\boldsymbol{L}^T\boldsymbol{L})^-$ (see Definition A.1). The integral is known irrespective of the type of normal approximation. Let k_i, $i = 1\ldots\kappa = \min(q,n)$ be the i^{th} non-zero eigenvalue of $\boldsymbol{K}^-$. Then

$$f(\boldsymbol{y}) \approx \frac{(2\pi\sigma^2)^{\kappa/2}}{(2\pi\sigma^2)^{n/2}(2\phi)^q}\left(\prod_{i=1}^{\kappa} k_i\right)^{1/2}\exp\left(h(\bar{\boldsymbol{\beta}},\sigma^2,\phi)\right).$$

The function $h(\bar{\boldsymbol{\beta}},\sigma^2,\phi)$ can be re-written by noticing that at the maximum of $h(\boldsymbol{\beta},\sigma^2,\phi)$ $\boldsymbol{L}^T\boldsymbol{y} = \boldsymbol{L}^T\boldsymbol{L}\bar{\boldsymbol{\beta}} + \frac{\sigma^2}{\phi}\bar{\boldsymbol{v}}$ and $\bar{\boldsymbol{v}}^T\bar{\boldsymbol{\beta}} = \|\bar{\boldsymbol{\beta}}\|_1$ so that

$$\begin{aligned}
h(\bar{\boldsymbol{\beta}},\sigma^2,\phi) &= -\frac{1}{2\sigma^2}\boldsymbol{y}^T\boldsymbol{y} + \frac{1}{\sigma^2}\bar{\boldsymbol{\beta}}^T\boldsymbol{L}^T\boldsymbol{y} - \frac{1}{2\sigma^2}\bar{\boldsymbol{\beta}}^T\boldsymbol{L}^T\boldsymbol{L}\bar{\boldsymbol{\beta}} - \frac{1}{\phi}\|\bar{\boldsymbol{\beta}}\|_1 \\
&= -\frac{1}{2\sigma^2}\boldsymbol{y}^T\boldsymbol{y} + \frac{1}{\sigma^2}\bar{\boldsymbol{\beta}}^T(\boldsymbol{L}^T\boldsymbol{L}\bar{\boldsymbol{\beta}} + \frac{\sigma^2}{\phi}\bar{\boldsymbol{v}}) - \frac{1}{2\sigma^2}\bar{\boldsymbol{\beta}}^T\boldsymbol{L}^T\boldsymbol{L}\bar{\boldsymbol{\beta}} - \frac{1}{\phi}\bar{\boldsymbol{v}}^T\bar{\boldsymbol{\beta}} \\
&= -\left(\frac{1}{2\sigma^2}\boldsymbol{y}^T\boldsymbol{y} - \frac{1}{2\sigma^2}\bar{\boldsymbol{\beta}}^T\boldsymbol{L}^T\boldsymbol{L}\bar{\boldsymbol{\beta}}\right).
\end{aligned}$$

The result is obtained by taking the logarithm of the marginal distribution and disregarding constant terms. □

7.3 Derivatives of LASSO Estimates

Optimisation of the likelihood requires the derivatives of the LASSO estimates with respect to the dispersion parameters. These are not obviously obtained as the estimates are themselves functions of the estimates $\bar{\boldsymbol{v}}$. Fortunately, there are some restrictions on the derivatives of $\bar{\boldsymbol{\beta}}$ and the corresponding $\bar{\boldsymbol{v}}$ that enable them to be calculated.

The derivation starts by partitioning the estimates into non-zero and zero estimates, $\bar{\boldsymbol{\beta}}_{\mathcal{S}}$ and $\bar{\boldsymbol{\beta}}_{\mathcal{Z}}$ respectively, and comformably partitioning $\bar{\boldsymbol{v}}$. Now, $\bar{\boldsymbol{\beta}}_{\mathcal{Z}}$ and $\bar{\boldsymbol{v}}_{\mathcal{S}}$ will be constant for a neighbourhood of the dispersion parameters. If they were not constant, then they would not be members of their respective sets. However, since they are constant, their derivatives will be zero. This argument forms the basis for calculating the derivatives of $\bar{\boldsymbol{\beta}}$ and hence optimising the likelihood ℓ.

The following Lemma is crucial for calculating the derivatives. The Theorem giving the derivatives will follow the Lemma directly.

Lemma 7.1. *Define the $q \times q$ matrix $\boldsymbol{D}$ as*

$$\boldsymbol{L}^T\boldsymbol{L} = \boldsymbol{D} = \begin{pmatrix} \boldsymbol{D}_{\mathcal{SS}} & \boldsymbol{D}_{\mathcal{SZ}} \\ \boldsymbol{D}_{\mathcal{ZS}} & \boldsymbol{D}_{\mathcal{ZZ}} \end{pmatrix} \quad \text{and} \quad (\boldsymbol{L}^T\boldsymbol{L})^- = \boldsymbol{D}^- = \begin{pmatrix} \boldsymbol{D}^{\mathcal{SS}} & \boldsymbol{D}^{\mathcal{SZ}} \\ \boldsymbol{D}^{\mathcal{ZS}} & \boldsymbol{D}^{\mathcal{ZZ}} \end{pmatrix}$$

where $\boldsymbol{L}$ is a $n \times q$ matrix and q may be larger than n, then

$$\left(\boldsymbol{I}_{\mathcal{S}}, -\boldsymbol{D}^{\mathcal{SZ}}\left(\boldsymbol{D}^{\mathcal{ZZ}}\right)^-\right)\boldsymbol{D}^- = \begin{pmatrix} \boldsymbol{D}^-_{\mathcal{SS}} & \boldsymbol{0} \end{pmatrix}.$$

Proof. Using Theorem B.1 the generalised inverse $\boldsymbol{D}^-$ can be written as

$$\boldsymbol{D}^- = \begin{pmatrix} \boldsymbol{D}^-_{\mathcal{SS}} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{pmatrix} + \begin{pmatrix} -\boldsymbol{D}^-_{\mathcal{SS}}\boldsymbol{D}_{\mathcal{SZ}} \\ \boldsymbol{I}_{\mathcal{Z}} \end{pmatrix} \boldsymbol{Q}^- \left(-\boldsymbol{D}_{\mathcal{ZS}}\boldsymbol{D}^-_{\mathcal{SS}}, \boldsymbol{I}_{\mathcal{S}}\right)$$

where $\boldsymbol{Q} = \boldsymbol{D}_{\mathcal{ZZ}} - \boldsymbol{D}_{\mathcal{ZS}}\left(\boldsymbol{D}_{\mathcal{SS}}\right)^- \boldsymbol{D}_{\mathcal{SZ}}$. Theorem B.2 shows that $\boldsymbol{Q}^- = \boldsymbol{D}^{\mathcal{ZZ}}$. To prove the Lemma it suffices to show that

$$\begin{aligned}
\left(\boldsymbol{I}_{\mathcal{S}}, -\boldsymbol{D}^{\mathcal{SZ}}\left(\boldsymbol{D}^{\mathcal{ZZ}}\right)^-\right)\begin{pmatrix} -\boldsymbol{D}^-_{\mathcal{SS}}\boldsymbol{D}_{\mathcal{SZ}} \\ \boldsymbol{I}_{\mathcal{Z}} \end{pmatrix}\boldsymbol{Q}^- &= -\left(\boldsymbol{D}^-_{\mathcal{SS}}\boldsymbol{D}_{\mathcal{SZ}} + \boldsymbol{D}^{\mathcal{SZ}}\left(\boldsymbol{D}^{\mathcal{ZZ}}\right)^-\right)\boldsymbol{D}^{\mathcal{ZZ}} \\
&= -\left(\boldsymbol{D}^-_{\mathcal{SS}}\boldsymbol{D}_{\mathcal{SZ}} - \boldsymbol{D}^-_{\mathcal{SS}}\boldsymbol{D}_{\mathcal{SZ}}\boldsymbol{D}^{\mathcal{ZZ}}\left(\boldsymbol{D}^{\mathcal{ZZ}}\right)^-\right)\boldsymbol{D}^{\mathcal{ZZ}} \\
&= -\boldsymbol{D}^-_{\mathcal{SS}}\boldsymbol{D}_{\mathcal{SZ}}\left(\boldsymbol{D}^{\mathcal{ZZ}} - \boldsymbol{D}^{\mathcal{ZZ}}\left(\boldsymbol{D}^{\mathcal{ZZ}}\right)^-\boldsymbol{D}^{\mathcal{ZZ}}\right) \\
&= \boldsymbol{0}
\end{aligned}$$

where the second line follows from using Theorem B.1 and the last line follows from the definition of a generalised inverse. □

Theorem 7.2. *The derivatives of the LASSO estimates with respect to σ^2 and ϕ are*

$$\begin{aligned}
\frac{\partial\bar{\boldsymbol{\beta}}}{\partial\sigma^2} &= -\frac{1}{\phi}\begin{pmatrix} (\boldsymbol{L}^T_{\mathcal{S}}\boldsymbol{L}_{\mathcal{S}})^{-1} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{pmatrix}\bar{\boldsymbol{v}} \qquad \text{and} \\
\frac{\partial\bar{\boldsymbol{\beta}}}{\partial\phi} &= -\frac{\sigma^2}{\phi^2}\begin{pmatrix} (\boldsymbol{L}^T_{\mathcal{S}}\boldsymbol{L}_{\mathcal{S}})^{-1} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{pmatrix}\bar{\boldsymbol{v}}.
\end{aligned}$$

Proof. The derivation for the derivative with respect to σ^2 is presented. The derivative with respect to ϕ follows similarly. The LASSO estimate must solve the equation $\partial_{\boldsymbol{\beta}} h(\boldsymbol{\beta}, \sigma^2, \phi) = \boldsymbol{0}$ and hence is given by

$$\bar{\boldsymbol{\beta}} \;=\; (\boldsymbol{L}^T\boldsymbol{L})^-\boldsymbol{L}^T\boldsymbol{y} - \frac{\sigma^2}{\phi}(\boldsymbol{L}^T\boldsymbol{L})^-\bar{\boldsymbol{v}} \;=\; \hat{\boldsymbol{\beta}} - \frac{\sigma^2}{\phi}\boldsymbol{D}^-\bar{\boldsymbol{v}}.$$

where $\boldsymbol{D}$ and $\boldsymbol{D}^-$ are defined as in Lemma 7.1. Recall that the derivatives of $\bar{\boldsymbol{\beta}}_{\mathcal{Z}}$ and of $\bar{\boldsymbol{v}}_{\mathcal{S}}$

are zero, so

$$
\begin{aligned}
\frac{\partial \bar{\boldsymbol{\beta}}_{\mathcal{Z}}}{\partial \sigma^2} &= \mathbf{0} \\
&= -\frac{1}{\phi}\boldsymbol{D}^{\mathcal{Z}\cdot}\bar{\boldsymbol{v}} - \frac{\sigma^2}{\phi}\boldsymbol{D}^{\mathcal{Z}\cdot}\frac{\partial \bar{\boldsymbol{v}}}{\partial \sigma^2} \\
&= -\frac{1}{\phi}\boldsymbol{D}^{\mathcal{Z}\cdot}\bar{\boldsymbol{v}} - \frac{\sigma^2}{\phi}\boldsymbol{D}^{\mathcal{Z}\mathcal{S}}\frac{\partial \bar{\boldsymbol{v}}_{\mathcal{S}}}{\partial \sigma^2} - \frac{\sigma^2}{\phi}\boldsymbol{D}^{\mathcal{Z}\mathcal{Z}}\frac{\partial \bar{\boldsymbol{v}}_{\mathcal{Z}}}{\partial \sigma^2} \\
&= -\frac{1}{\phi}\boldsymbol{D}^{\mathcal{Z}\cdot}\bar{\boldsymbol{v}} - \frac{\sigma^2}{\phi}\boldsymbol{D}^{\mathcal{Z}\mathcal{Z}}\frac{\partial \bar{\boldsymbol{v}}_{\mathcal{Z}}}{\partial \sigma^2} \\
\therefore \qquad \frac{\partial \bar{\boldsymbol{v}}_{\mathcal{Z}}}{\partial \sigma^2} &= -\frac{1}{\sigma^2}\left(\boldsymbol{D}^{\mathcal{Z}\mathcal{Z}}\right)^{-}\boldsymbol{D}^{\mathcal{Z}\cdot}\bar{\boldsymbol{v}}.
\end{aligned}
$$

Now,

$$
\begin{aligned}
\frac{\partial \bar{\boldsymbol{\beta}}_{\mathcal{S}}}{\partial \sigma^2} &= \frac{1}{\phi^2}\boldsymbol{D}^{\mathcal{S}\cdot}\bar{\boldsymbol{v}} - \frac{\sigma^2}{\phi}\boldsymbol{D}^{\mathcal{S}\cdot}\frac{\partial \bar{\boldsymbol{v}}}{\partial \sigma^2} \\
&= \frac{1}{\phi^2}\boldsymbol{D}^{\mathcal{S}\cdot}\bar{\boldsymbol{v}} - \frac{\sigma^2}{\phi}\boldsymbol{D}^{\mathcal{S}\mathcal{S}}\frac{\partial \bar{\boldsymbol{v}}_{\mathcal{S}}}{\partial \sigma^2} - \frac{\sigma^2}{\phi}\boldsymbol{D}^{\mathcal{S}\mathcal{Z}}\frac{\partial \bar{\boldsymbol{v}}_{\mathcal{Z}}}{\partial \sigma^2} \\
&= \frac{1}{\phi^2}\boldsymbol{D}^{\mathcal{S}\cdot}\bar{\boldsymbol{v}} - \frac{\sigma^2}{\phi}\boldsymbol{D}^{\mathcal{S}\mathcal{Z}}\frac{\partial \bar{\boldsymbol{v}}_{\mathcal{Z}}}{\partial \sigma^2} \\
&= \frac{1}{\phi^2}\boldsymbol{D}^{\mathcal{S}\cdot}\bar{\boldsymbol{v}} + \frac{1}{\phi}\boldsymbol{D}^{\mathcal{S}\mathcal{Z}}\left(\boldsymbol{D}^{\mathcal{Z}\mathcal{Z}}\right)^{-}\boldsymbol{D}^{\mathcal{S}\cdot}\bar{\boldsymbol{v}} \\
&= -\frac{1}{\phi}\left[\boldsymbol{I}_{\mathcal{S}}, -\boldsymbol{D}^{\mathcal{S}\mathcal{Z}}\left(\boldsymbol{D}^{\mathcal{Z}\mathcal{Z}}\right)^{-}\right]\boldsymbol{D}^{-}\bar{\boldsymbol{v}} \\
&= -\frac{1}{\phi}\begin{pmatrix} \boldsymbol{D}_{\mathcal{S}\mathcal{S}}^{-} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}\bar{\boldsymbol{v}} \qquad \text{by Lemma 7.1.}
\end{aligned}
$$

Theorem 6 of Osborne et al. (2000) states that there will be at most n elements in $\boldsymbol{\mathcal{S}}$. This means that the matrix $\boldsymbol{D}_{\mathcal{S}\mathcal{S}} = \boldsymbol{L}_{\mathcal{S}}^T\boldsymbol{L}_{\mathcal{S}}$ will be full rank and uniquely invertible. That is $\boldsymbol{D}_{\mathcal{S}\mathcal{S}}^{-} = \boldsymbol{D}_{\mathcal{S}\mathcal{S}}^{-1}$. □

7.4 Score Equations and Estimates

It is now possible to specify the score equations (first derivatives of the log-likelihood in Theorem 7.1). When assigned to zero and solved, these provide the estimating equations.

Theorem 7.3. *The score equations for the dispersion parameters σ^2 and ϕ are*

$$
\begin{aligned}
\frac{\partial \ell_r(\sigma^2, \phi)}{\partial \sigma^2} &= -\frac{(n-\kappa)}{2\sigma^2} + \frac{1}{2\sigma^4}\left(\boldsymbol{y}^T\boldsymbol{y} - \bar{\boldsymbol{\beta}}^T\boldsymbol{L}^T\boldsymbol{L}\bar{\boldsymbol{\beta}}\right) - \frac{1}{\sigma^2\phi}\|\bar{\boldsymbol{\beta}}\|_1 \\
\frac{\partial \ell_r(\sigma^2, \phi)}{\partial \phi} &= -\frac{q}{\phi} + \frac{1}{\phi^2}\|\bar{\boldsymbol{\beta}}\|_1.
\end{aligned}
$$

Proof. The score for the residual variance σ^2, using the derivative of the LASSO estimates in Theorem 7.2, is

$$\begin{aligned}\frac{\partial \ell_r(\sigma^2,\phi)}{\partial \sigma^2} &= -\frac{(n-\kappa)}{2\sigma^2} + \frac{1}{2\sigma^4}\left(\boldsymbol{y}^T\boldsymbol{y} - \bar{\boldsymbol{\beta}}^T\boldsymbol{L}^T\boldsymbol{L}\bar{\boldsymbol{\beta}}\right) - \frac{1}{\sigma^2\phi}\bar{\boldsymbol{v}}^T\begin{pmatrix}\left(\boldsymbol{L}_{\mathcal{S}}^T\boldsymbol{L}_{\mathcal{S}}\right)^{-1} & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{0}\end{pmatrix}\boldsymbol{L}^T\boldsymbol{L}\bar{\boldsymbol{\beta}}\\ &= -\frac{(n-\kappa)}{2\sigma^2} + \frac{1}{2\sigma^4}\left(\boldsymbol{y}^T\boldsymbol{y} - \bar{\boldsymbol{\beta}}^T\boldsymbol{L}^T\boldsymbol{L}\bar{\boldsymbol{\beta}}\right) - \frac{1}{\sigma^2\phi}\bar{\boldsymbol{v}}_{\mathcal{S}}^T\bar{\boldsymbol{\beta}}_{\mathcal{S}}\\ &= -\frac{(n-\kappa)}{2\sigma^2} + \frac{1}{2\sigma^4}\left(\boldsymbol{y}^T\boldsymbol{y} - \bar{\boldsymbol{\beta}}^T\boldsymbol{L}^T\boldsymbol{L}\bar{\boldsymbol{\beta}}\right) - \frac{1}{\sigma^2\phi}\|\bar{\boldsymbol{\beta}}\|_1.\end{aligned}$$

The score for the dispersion parameter ϕ is

$$\begin{aligned}\frac{\partial \ell_r(\sigma^2,\phi)}{\partial \phi} &= -\frac{q}{\phi} + \frac{1}{\phi^2}\bar{\boldsymbol{v}}^T\begin{pmatrix}(\boldsymbol{L}_{\mathcal{S}}^T\boldsymbol{L}_{\mathcal{S}})^{-1} & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{0}\end{pmatrix}\boldsymbol{L}^T\boldsymbol{L}\bar{\boldsymbol{\beta}}\\ &= -\frac{q}{\phi} + \frac{1}{\phi^2}\bar{\boldsymbol{v}}_{\mathcal{S}}^T\bar{\boldsymbol{\beta}}_{\mathcal{S}}\\ &= -\frac{q}{\phi} + \frac{1}{\phi^2}\|\bar{\boldsymbol{\beta}}\|_1.\end{aligned}$$

□

Assigning the score equations to be zero and solving them gives estimates for the dispersion parameters. These have closed forms and are given in the following Theorem.

Theorem 7.4. *The estimates, for given $\bar{\boldsymbol{\beta}}$, for the dispersion parameters σ^2 and ϕ are*

$$\begin{aligned}\hat{\sigma}^2 &= \frac{\left(\boldsymbol{y}^T\boldsymbol{y} - \bar{\boldsymbol{\beta}}^T\boldsymbol{L}^T\boldsymbol{L}\bar{\boldsymbol{\beta}}\right)}{n-\kappa+\frac{2}{\phi}\|\bar{\boldsymbol{\beta}}\|_1}\\ \hat{\phi} &= \frac{\|\bar{\boldsymbol{\beta}}\|_1}{q}.\end{aligned}$$

Proof. Consider the estimate for σ^2

$$\begin{aligned}\frac{\partial \ell_r(\sigma^2,\phi)}{\partial \sigma^2} &= 0\\ &= -\frac{1}{2\hat{\sigma}^2}\left(\frac{(n-\kappa)}{2} + \frac{2}{\phi}\|\bar{\boldsymbol{\beta}}\|_1\right) + \frac{1}{2\hat{\sigma}^4}\left(\boldsymbol{y}^T\boldsymbol{y} - \bar{\boldsymbol{\beta}}^T\boldsymbol{L}^T\boldsymbol{L}\bar{\boldsymbol{\beta}}\right)\\ \therefore \qquad \hat{\sigma}^2\left(n-\kappa+\frac{2}{\phi}\|\bar{\boldsymbol{\beta}}\|_1\right) &= \left(\boldsymbol{y}^T\boldsymbol{y} - \bar{\boldsymbol{\beta}}^T\boldsymbol{L}^T\boldsymbol{L}\bar{\boldsymbol{\beta}}\right)\\ \therefore \qquad \hat{\sigma}^2 &= \frac{\left(\boldsymbol{y}^T\boldsymbol{y} - \bar{\boldsymbol{\beta}}^T\boldsymbol{L}^T\boldsymbol{L}\bar{\boldsymbol{\beta}}\right)}{n-\kappa+\frac{2}{\phi}\|\bar{\boldsymbol{\beta}}\|_1}.\end{aligned}$$

Now, consider the estimate for ϕ

$$\begin{aligned} \frac{\partial \ell_r(\sigma^2, \phi)}{\partial \phi} &= 0 \\ &= -\frac{q}{\hat{\phi}} + \frac{1}{\hat{\phi}^2}\|\bar{\boldsymbol{\beta}}\|_1 \\ \therefore \qquad \hat{\phi} q &= \|\bar{\boldsymbol{\beta}}\|_1 \\ \therefore \qquad \hat{\phi} &= \frac{\|\bar{\boldsymbol{\beta}}\|_1}{q}. \end{aligned}$$

□

7.4.1 *Bias in Score Equations and Estimates*

The unbiasedness of the estimates obtained from using Theorem 7.4 is investigated using a simulation study. The simulation study was performed by generating $n = 100$ outcomes from the model (7.1.1), using a common 100×30 design matrix. Three distributions of the $q = 30$ LASSO random effects were used, all with mean zero and varying variances. Three distributions for the residuals, assumed to be independently and identically distributed, were also used. The bias in both the score equations and the resulting estimates were calculated at each combination of LASSO effect variance and residual variance. The levels of the dispersion parameters were chosen to provide a spread in the number of non-zero LASSO estimates. The results are presented in Table 7.1.

Table 7.1: Average score and dispersion parameters for 1000 simulations of model (7.1.1) at chosen combinations of σ^2 and ϕ. The scores are calculated using Theorem 7.3 and the estimates calculated using Theorem 7.4.

		$\hat{\sigma}^{2\dagger}$			$\hat{\phi}^{\ddagger}$		
		$\phi = 0.05$	$\phi = 0.3$	$\phi = 1$	$\phi = 0.05$	$\phi = 0.3$	$\phi = 1$
Score	$\sigma^2 = 0.25$	106.38	118.07	118.87	-282.77	-2.20	0.02
	$\sigma^2 = 1$	21.05	29.61	29.85	-529.06	-8.84	-0.21
	$\sigma^2 = 2$	9.27	14.42	15.12	-589.16	-15.64	-0.21
Estimate	$\sigma^2 = 0.25$	0.30	0.25	0.25	2.64×10^{-2}	0.29	1.00
	$\sigma^2 = 1$	1.45	1.04	1.01	5.91×10^{-3}	0.27	0.99
	$\sigma^2 = 2$	3.01	2.12	2.01	9.03×10^{-4}	0.25	0.99

† Standard errors for scores: < 1.09 for $\sigma^2 = 0.25$, < 0.26 for $\sigma^2 = 1$, < 0.13 for $\sigma^2 = 2$ and for estimates: < 0.002 for $\sigma^2 = 0.25$, < 0.006 for $\sigma^2 = 1$, < 0.014 for $\sigma^2 = 2$.
‡ Standard errors for scores: < 3.17 for $\phi = 0.05$, < 0.62 for $\phi = 0.3$, < 0.18 for $\phi = 1$ and for estimates: < 0.001 for $\phi = 0.05$, < 0.002 for $\phi = 0.3$ and < 0.006 for $\phi = 1$.

If the score equations are unbiased, then their expectation and their empirical means should be zero. However this is not the case (Table 7.1). This also leads to bias in the dispersion parameter estimates. The level of bias is greatest for large values of the penalty parameter σ^2/ϕ. In these situations the number of non-zero LASSO estimates is small, for the $\sigma^2 = 2$ and $\phi = 0.05$ simulation the average number of non-zero estimates was only 0.02 non-zero LASSO estimates per data set.

The bias can be substantially alleviated, but not totally removed, by replacing the number of LASSO effects q by the number of non-zero LASSO estimates $q_{\mathcal{S}}$ in the score equations and the resulting dispersion parameter estimates. However, this is an ad-hoc procedure. The results of the simulation, over the same simulated data sets as the simulation for the un-modified scores and estimates, are given in Table 7.2. Compared with Table 7.1, there is less bias. However, the bias is not totally removed. Thus, there is still a need for a better method to adjust the score equations. A method to perform this is presented in the next section.

Table 7.2: Average score and dispersion parameters for 1000 simulations of model (7.1.1) at chosen combinations of σ^2 and ϕ. The scores and estimates are calculated by replacing q with $q_{\mathcal{S}}$ in Theorems 7.3 and 7.4.

		$\hat{\sigma}^{2\dagger}$			$\hat{\phi}^{\ddagger}$		
		$\phi = 0.05$	$\phi = 0.3$	$\phi = 1$	$\phi = 0.05$	$\phi = 0.3$	$\phi = 1$
	$\sigma^2 = 0.25$	74.76	116.11	118.67	33.37	1.08	0.12
Score	$\sigma^2 = 1$	7.57	27.90	29.65	10.04	2.52	0.21
	$\sigma^2 = 2$	1.89	12.92	14.93	1.46	4.38	0.56
	$\sigma^2 = 0.25$	0.26	0.25	0.25	0.056	0.30	1.00
Estimate	$\sigma^2 = 1$	1.08	1.01	1.00	0.056	0.31	1.01
	$\sigma^2 = 2$	2.13	2.02	2.00	0.021	0.32	1.02

† Standard errors for scores: < 1.18 for $\sigma^2 = 0.25$, < 0.27 for $\sigma^2 = 1$, < 0.14 for $\sigma^2 = 2$ and for estimates: < 0.001 for $\sigma^2 = 0.25$, < 0.005 for $\sigma^2 = 1$, < 0.01 for $\sigma^2 = 2$.
‡ Standard errors for scores: < 2.53 for $\phi = 0.05$, < 0.58 for $\phi = 0.3$, < 0.18 for $\phi = 1$ and for estimates: < 0.002 for $\phi = 0.05$, < 0.002 for $\phi = 0.3 < 0.006$ for $\phi = 1$.

7.5 Adjusted Score Equations

A method of adjusting the scores and information for estimation of parameters when nuisance parameters are present was specified in McCullagh & Tibshirani (1990). The score equations are adjusted so that they are unbiased, that is they have zero expectation. The score equations in Theorem 7.3 do not contain nuisance parameters but the method of McCullagh & Tibshirani (1990) can still be used. That is, the score equations can be adjusted so that they have zero expectation.

Let the expected values of the score for σ^2 and ϕ be M_{σ^2} and M_ϕ respectively. The estimates obtained from the adjusted score equations are given in the following Theorem.

Theorem 7.5. *The estimates for σ^2 and ϕ arising from the mean adjusted score equations are*

$$\begin{aligned}\hat{\sigma}^2 &= \frac{\boldsymbol{y}^T\boldsymbol{y} - \bar{\boldsymbol{\beta}}^T\boldsymbol{L}^T\boldsymbol{L}\bar{\boldsymbol{\beta}}}{n - \kappa + \frac{2}{\phi}\|\bar{\boldsymbol{\beta}}\|_1 + 2\hat{\sigma}^2 M_{\sigma^2}} \qquad \textit{and} \\ \hat{\phi} &= \frac{\|\bar{\boldsymbol{\beta}}\|_1}{q + \hat{\phi}M_\phi}.\end{aligned}$$

Proof. Firstly, consider σ^2. The estimates are those that solve the equation

$$\begin{aligned} & -\frac{(n-\kappa)}{2\hat{\sigma}^2} + \frac{1}{2\hat{\sigma}^4}\left(\boldsymbol{y}^T\boldsymbol{y} - \bar{\boldsymbol{\beta}}^T\boldsymbol{L}^T\boldsymbol{L}\bar{\boldsymbol{\beta}}\right) - \frac{1}{\hat{\sigma}^2\phi}\|\bar{\boldsymbol{\beta}}\|_1 - M_{\sigma^2} = 0 \\ \therefore \quad & \frac{1}{2\hat{\sigma}^2}\left(n - \kappa + \frac{2}{\phi}\|\bar{\boldsymbol{\beta}}\|_1 + 2\hat{\sigma}^2 M_{\sigma^2}\right) = \frac{1}{2\hat{\sigma}^4}\left(\boldsymbol{y}^T\boldsymbol{y} - \bar{\boldsymbol{\beta}}^T\boldsymbol{L}^T\boldsymbol{L}\bar{\boldsymbol{\beta}}\right) \\ \therefore \quad & \hat{\sigma}^2 = \frac{\boldsymbol{y}^T\boldsymbol{y} - \bar{\boldsymbol{\beta}}^T\boldsymbol{L}^T\boldsymbol{L}\bar{\boldsymbol{\beta}}}{n - \kappa + \frac{2}{\phi}\|\bar{\boldsymbol{\beta}}\|_1 + 2\hat{\sigma}^2 M_{\sigma^2}}.\end{aligned}$$

Now consider ϕ.

$$\begin{aligned} & -\frac{q}{\hat{\phi}} + \frac{1}{\hat{\phi}^2}\|\bar{\boldsymbol{\beta}}\|_1 - M_\phi = 0 \\ \therefore \quad & \hat{\phi}(q + \hat{\phi}M_\phi) = \|\bar{\boldsymbol{\beta}}\|_1 \\ \therefore \quad & \hat{\phi} = \frac{\|\bar{\boldsymbol{\beta}}\|_1}{q + \hat{\phi}M_\phi}.\end{aligned}$$

□

The estimates in Theorem 7.5 arise from a quadratic equation. They are not solved as a quadratic, rather the form of the original estimates in Theorem 7.4 is maintained. This form makes the adjustment to the denominator obvious. Terms in the denominator of both estimates are essentially the degrees of freedom for the LASSO effects. The estimates in this form require iteration to calculate, which is already necessary due to the presence of the LASSO estimates $\bar{\boldsymbol{\beta}}$.

7.5.1 *Theoretical Expected Scores*

Exact theoretical calculation of the expected values of the score equations is not possible as the first and second moments of the distribution of the LASSO estimates and the corresponding estimates of $\bar{\boldsymbol{v}}$ are unknown.

It may be possible to obtain a reasonable approximation to the expected scores or their constituent components by using the substitution $\mathrm{E}\left(\bar{\boldsymbol{v}}\right) = \bar{\boldsymbol{v}}$. That is, using the observed $\bar{\boldsymbol{v}}$ in place of its expected value. It turns out that this approximation is not sensible. However, the investigation is presented for completeness.

Theorem 7.6. *With the assumption that* $\mathrm{E}\left(\bar{\boldsymbol{v}}\right) = \bar{\boldsymbol{v}}$ *then*

$$\begin{aligned}
\mathrm{E}\left(\bar{\boldsymbol{\beta}}^T \boldsymbol{L}^T \boldsymbol{L} \bar{\boldsymbol{\beta}}\right) &= \mathrm{tr}\left(\boldsymbol{L}^T \boldsymbol{L}\left(2\phi^2 \boldsymbol{I}_q - \sigma^2 (\boldsymbol{L}^T \boldsymbol{L})^-\right)\right) + \frac{\sigma^4}{\phi^2} \bar{\boldsymbol{v}}^T (\boldsymbol{L}^T \boldsymbol{L})^- \bar{\boldsymbol{v}} \qquad \textit{and} \\
\mathrm{E}\left(\|\bar{\boldsymbol{\beta}}\|_1\right) &= -\frac{\sigma^2}{\phi} \bar{\boldsymbol{v}}^T (\boldsymbol{L}^T \boldsymbol{L})^- \bar{\boldsymbol{v}}.
\end{aligned}$$

Proof. The expected value and variance of $\hat{\boldsymbol{\beta}} = (\boldsymbol{L}^T \boldsymbol{L})^- \boldsymbol{L}^T \boldsymbol{y}$, along with the result for expectations for quadratic forms (Theorem A.4) is needed.

$$\begin{aligned}
\mathrm{E}\left(\boldsymbol{y}\right) &= \boldsymbol{0} \\
\mathrm{var}\left(\boldsymbol{y}\right) &= 2\phi^2 \boldsymbol{L}\boldsymbol{L}^T + \sigma^2 \boldsymbol{I}_n \\
\mathrm{E}\left(\hat{\boldsymbol{\beta}}\right) &= \mathrm{E}\left((\boldsymbol{L}^T \boldsymbol{L})^- \boldsymbol{L}^T \boldsymbol{y}\right) \\
&= \boldsymbol{0} \\
\mathrm{var}\left(\hat{\boldsymbol{\beta}}\right) &= (\boldsymbol{L}^T \boldsymbol{L})^- \boldsymbol{L}^T \mathrm{var}\left(\boldsymbol{y}\right) \boldsymbol{L} (\boldsymbol{L}^T \boldsymbol{L})^- \\
&= 2\phi^2 (\boldsymbol{L}^T \boldsymbol{L})^- \boldsymbol{L}^T \boldsymbol{L}\boldsymbol{L}^T \boldsymbol{L} (\boldsymbol{L}^T \boldsymbol{L})^- + \sigma^2 (\boldsymbol{L}^T \boldsymbol{L})^- \boldsymbol{L}^T \boldsymbol{L} (\boldsymbol{L}^T \boldsymbol{L})^- \\
&= (\boldsymbol{L}^T \boldsymbol{L})^- \boldsymbol{L}^T \boldsymbol{L}\left(2\phi^2 \boldsymbol{L}^T \boldsymbol{L} + \sigma^2 \boldsymbol{I}_q\right)(\boldsymbol{L}^T \boldsymbol{L})^-.
\end{aligned}$$

The expected value of the quadratic form $\hat{\boldsymbol{\beta}}^T \boldsymbol{L}^T \boldsymbol{L} \hat{\boldsymbol{\beta}}$ can now be evaluated. Using Theorem A.4, it is

$$\begin{aligned}
\hat{\boldsymbol{\beta}}^T \boldsymbol{L}^T \boldsymbol{L} \hat{\boldsymbol{\beta}} &= \mathrm{tr}\left(\boldsymbol{L}^T \boldsymbol{L} (\boldsymbol{L}^T \boldsymbol{L})^- \boldsymbol{L}^T \boldsymbol{L}\left(2\phi^2 \boldsymbol{L}^T \boldsymbol{L} + \sigma^2 \boldsymbol{I}_q\right)(\boldsymbol{L}^T \boldsymbol{L})^-\right) \\
&= \mathrm{tr}\left(\left(2\phi^2 \boldsymbol{L}^T \boldsymbol{L} + \sigma^2 \boldsymbol{I}_q\right)(\boldsymbol{L}^T \boldsymbol{L})^- \boldsymbol{L}^T \boldsymbol{L}\right) \\
&= \mathrm{tr}\left(\left(2\phi^2 \boldsymbol{I}_q + \sigma^2 (\boldsymbol{L}^T \boldsymbol{L})^-\right) \boldsymbol{L}^T \boldsymbol{L}\right).
\end{aligned}$$

Consider $\mathrm{E}\left(\bar{\boldsymbol{\beta}}^T \boldsymbol{L}^T \boldsymbol{L} \bar{\boldsymbol{\beta}}\right)$ first.

$$\begin{aligned}
\mathrm{E}\left(\bar{\boldsymbol{\beta}}^T \boldsymbol{L}^T \boldsymbol{L} \bar{\boldsymbol{\beta}}\right) &= \mathrm{E}\left(\hat{\boldsymbol{\beta}}^T \boldsymbol{L}^T \boldsymbol{L} \hat{\boldsymbol{\beta}} - 2\frac{\sigma^2}{\phi} \bar{\boldsymbol{v}}^T (\boldsymbol{L}^T \boldsymbol{L})^- \boldsymbol{L}^T \boldsymbol{L} \hat{\boldsymbol{\beta}} + \frac{\sigma^4}{\phi^2} \bar{\boldsymbol{v}}^T (\boldsymbol{L}^T \boldsymbol{L})^- \bar{\boldsymbol{v}}\right) \\
&= \mathrm{E}\left(\hat{\boldsymbol{\beta}}^T \boldsymbol{L}^T \boldsymbol{L} \hat{\boldsymbol{\beta}}\right) - 2\frac{\sigma^2}{\phi} \bar{\boldsymbol{v}}^T (\boldsymbol{L}^T \boldsymbol{L})^- \boldsymbol{L}^T \boldsymbol{L} \mathrm{E}\left(\hat{\boldsymbol{\beta}}\right) + \frac{\sigma^4}{\phi^2} \bar{\boldsymbol{v}}^T (\boldsymbol{L}^T \boldsymbol{L})^- \bar{\boldsymbol{v}} \\
&= \mathrm{tr}\left(\left(2\phi^2 \boldsymbol{I}_q + \sigma^2 (\boldsymbol{L}^T \boldsymbol{L})^-\right) \boldsymbol{L}^T \boldsymbol{L}\right) + \frac{\sigma^4}{\phi^2} \bar{\boldsymbol{v}}^T (\boldsymbol{L}^T \boldsymbol{L})^- \bar{\boldsymbol{v}}.
\end{aligned}$$

Now consider $\mathrm{E}\left(\|\bar{\boldsymbol{\beta}}\|_1\right) = \mathrm{E}\left(\bar{\boldsymbol{v}}^T \bar{\boldsymbol{\beta}}\right)$.

$$\begin{aligned}
\mathrm{E}\left(\bar{\boldsymbol{v}}^T \bar{\boldsymbol{\beta}}\right) &= \bar{\boldsymbol{v}}^T \mathrm{E}\left(\hat{\boldsymbol{\beta}} - \frac{\sigma^2}{\phi} (\boldsymbol{L}^T \boldsymbol{L})^- \bar{\boldsymbol{v}}\right) \\
&= -\frac{\sigma^2}{\phi} \bar{\boldsymbol{v}}^T (\boldsymbol{L}^T \boldsymbol{L})^- \bar{\boldsymbol{v}}.
\end{aligned}$$

□

There is a problem with the expected values in Theorem 7.6. The expected value of $\|\bar{\boldsymbol{\beta}}\|_1$ is negative as the quadratic form involving $\bar{\boldsymbol{v}}$ is always positive. This assumption obviously leads to an inappropriate expected score. Nevertheless, the performance of these expected scores is now assessed via a small simulation study.

The simulation study consisted of simulating 1000 data sets from the model (7.1.1) with $n = 100$ and $q = 20$. The design matrix $\boldsymbol{L}$ was generated once and held constant throughout the 1000 simulations. The dispersion values used to simulate the data were $\sigma^2 = 1$ and $\phi = 0.274$. The dispersion parameters used to calculate the score equations were $\sigma_*^2 = 1.5$ and $\phi_* = 0.245$. These were chosen to be different from the simulation values because non-zero scores were desirable. Also, the true values are never known prior to estimation.

For each simulation the values of $\bar{\boldsymbol{\beta}}^T \boldsymbol{L}\boldsymbol{L}\bar{\boldsymbol{\beta}}$, $\|\bar{\boldsymbol{\beta}}\|_1$, their expectations and both scores were recorded. If the expected values represent their corresponding observations, then the mean of the observed quantities and the expectations should be equal. Further, the expected values should have smaller variance. Results are given in Figure 7.1 and Table 7.3. The expectations are not constant as the expectations are functions of the stochastic variable $\bar{\boldsymbol{v}}$.

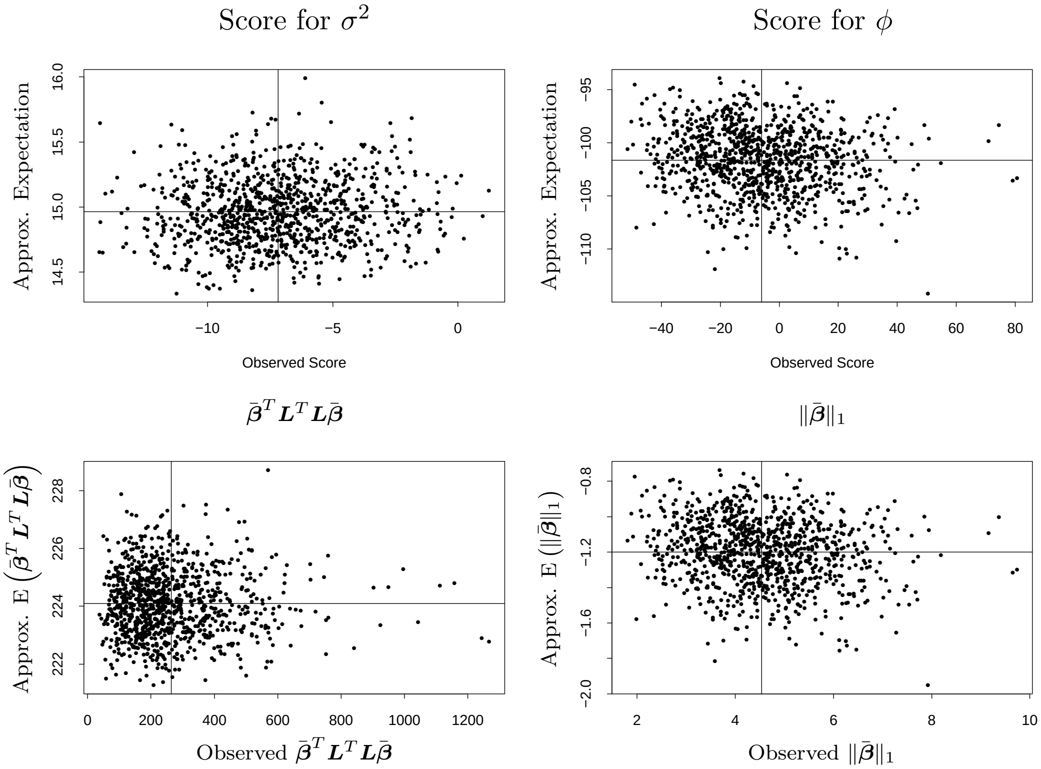

Figure 7.1: Plots of approximate analytical values versus observed values. The vertical (horizontal) lines are the means of the observed (expected) values.

The simulation results show that the expected score cannot be usefully calculated by assuming that $\mathrm{E}(\bar{\boldsymbol{v}}) = \bar{\boldsymbol{v}}$. The expected scores are not only substantially biased, but are also of the wrong sign for $\partial \ell_r / \partial \sigma^2$. Hence, adjusting the observed score with this expectation

Table 7.3: Means and variances of observed values and their expectations over 1000 simulations.

	Observed		Expected	
	Mean	Variance	Mean	Variance
$\partial\ell_r/\partial\sigma^2$	-7.18	7.56	14.97	0.06
$\partial\ell_r/\partial\phi$	-6.01	407.68	-101.63	9.64
$\bar{\boldsymbol{\beta}}^T\boldsymbol{L}^T\boldsymbol{L}\bar{\boldsymbol{\beta}}$	263.59	24826.64	224.10	1.30
$\lVert\bar{\boldsymbol{\beta}}\rVert_1$	4.54	1.47	-1.2	0.03

will exacerbate the problem rather than alleviate it!

The expected scores do have substantially less variation than the observed scores. This implies that the majority of the variation in $\bar{\boldsymbol{\beta}}$ comes from $\hat{\boldsymbol{\beta}}$ and not from $\frac{\sigma^2}{\phi}(\boldsymbol{L}^T\boldsymbol{L})^{-}\bar{\boldsymbol{v}}$. This is not totally unexpected as the elements in $\bar{\boldsymbol{v}}$ are bounded by -1 and 1. The amount of variance reduction will depend on the relative values of σ^2 and ϕ. The larger the ratio σ^2/ϕ, the less variance reduction. However, since the resulting expected score equations are biased this observation has not been investigated further.

7.5.2 *Empirical Expected Scores*

An alternative to calculating the theoretical expected score is to calculate an empirical version. Here, a parametric bootstrap (e.g. Davison & Hinkley, 1997) is used. This approach is similar to the bootstrap adjustment term used by McCullagh & Tibshirani (1990).

The scores are to be calculated at a specific value of σ^2 and ϕ. The bootstrap expectation is found from repeatedly simulating from the random effects model (7.1.1) using the given values of σ^2 and ϕ. For each data set the scores are calculated. The average of the scores for each data set are the bootstrap estimates of the expected scores. If the observed scores are highly variable then a larger number of bootstrap samples will be needed to obtain a specified level of precision in the expected scores. Unfortunately, the scores can be highly variable (Table 7.3).

Under the same simulation setup used to highlight the bias in the score equations in Section 7.4.1, the bootstrap adjusted scores and subsequent estimates were assessed. The same data sets as those in the simulations in Section 7.4.1 were used so the results are directly comparable. The bootstrap expectation was calculated using 200 bootstrap samples prior to simulating the 1000 data sets for calculation of the scores. The resulting score equations and estimates are presented in Table 7.4.

Comparing Table 7.4 to Tables 7.1 and 7.2, it is obvious that the bootstrap adjusted scores and the subsequent estimates are much less biased than the unadjusted and the q_S version of the unadjusted. This estimation comes at a cost in that the computation required to calculate the estimates is much greater.

Table 7.4: Average score and dispersion parameters for 1000 simulations of model (7.1.1) at chosen combinations of σ^2 and ϕ. The scores are adjusted by subtracting a bootstrap expected value using 200 bootstrap samples. The estimates are calculated by the results in 7.5.

		$\hat{\sigma}^{2\dagger}$			$\hat{\phi}^{\ddagger}$		
		$\phi = 0.05$	$\phi = 0.3$	$\phi = 1$	$\phi = 0.05$	$\phi = 0.3$	$\phi = 1$
	$\sigma^2 = 0.25$	-0.45	-0.54	2.12	-3.77	0.74	0.24
Score	$\sigma^2 = 1$	-0.91	1.25	-0.93	-1.18	1.11	-0.52
	$\sigma^2 = 2$	-0.26	0.09	-0.02	0.92	0.53	0.37
	$\sigma^2 = 0.25$	0.25	0.25	0.25	0.049	0.30	1.01
Estimate	$\sigma^2 = 1$	0.98	1.02	0.99	0.049	0.30	0.98
	$\sigma^2 = 2$	1.98	2.01	2.00	0.055	0.30	1.01

† Standard errors for scores: < 1.09 for $\sigma^2 = 0.25$, < 0.26 for $\sigma^2 = 1$, < 0.13 for $\sigma^2 = 2$ and for estimates: < 0.002 for $\sigma^2 = 0.25$, < 0.005 for $\sigma^2 = 1$, < 0.01 for $\sigma^2 = 2$.
‡ Standard errors for scores: < 3.16 for $\phi = 0.05$, < 0.62 for $\phi = 0.3$, < 0.18 for $\phi = 1$ and for estimates: < 0.004 for $\phi = 0.05$, < 0.003 for $\phi = 0.3$ < 0.006 for $\phi = 1$.

7.6 Estimation Algorithm Overview

All the tools necessary to estimate the dispersion parameters have now been developed. Estimation proceeds via iteration. Firstly, dispersion parameter estimates are assigned initial values so that there are non-zero LASSO estimates. This is achieved by finding the maximum ratio $\lambda = \sigma^2/\phi$ and choosing the initial ratio to be 10% of that value. Iteration now starts. The LASSO estimates are calculated using the last iteration's dispersion parameter estimates. The previous iteration's dispersion parameters are then used to calculate the score adjustment terms in Section 7.5.2. The score adjustment terms and the LASSO estimates are then used to update the dispersion parameters according to Theorem 7.5. This process of three steps is iterated over until convergence of the dispersion parameter estimates to some predefined tolerance level is reached.

The use of the bootstrap introduces a stochastic element into the estimation routine. This means that convergence can never be guaranteed. This can be overcome by taking partial steps from the previous estimate in the direction of the new estimate. This procedure is similar to the step-halving procedure (e.g. Jennrich & Sampson, 1976). When maximising likelihoods for variance components, Jennrich & Sampson (1976) used an iteratively reduced step size to avoid overshooting the maxima. If $\hat{\phi}^{(i+1)}$ is the updated estimate from Theorem 7.5 then the update chosen is

$$\hat{\phi}_*^{(i+1)} = \hat{\phi}^{(i)} + \nu \left(\hat{\phi}^{(i+1)} - \hat{\phi}^{(i)} \right)$$

where $0 < \nu < 1$ and is reduced from one iteration to the next. The reduction is implemented by multiplying ν by a chosen constant between 0 and 1. In practice, a value close to 1 is recommended to prevent inflating the risk of not finding the solution to the adjusted scores.

At each iteration the updated $\hat{\phi}$ may be zero. This occurs when $\bar{\boldsymbol{\beta}} = \mathbf{0}$, at which point all subsequent estimates of ϕ will also be zero. This may not be warranted and can be guarded against algorithmically. In this implementation an updated $\hat{\phi}$ that is zero is replaced by a fraction (one quarter) of its previous value. The estimate may be zero and this possibility is allowed for by limiting the number of times a zero estimate is replaced by the fraction of the previous. In this implementation the replacement is allowed for 10 times (default).

If the method of adjusting the score equations in Theorem 7.3 by replacing q by $q_{\mathcal{S}}$ is used, then care in estimation also needs to be taken. Difficulties potentially arise as the estimates from one iteration to the next tend to cycle between different neighbourhoods of estimate values. To avoid this occurring, the step reduction technique can once again be employed.

7.6.1 *Number of Bootstrap Samples*

With all bootstrap methods the number of bootstrap samples must be specified. If too few samples are used, then the expected values could be highly variable and the resulting estimates may not be accurate. If too many samples are used then the computational burden will be great.

In the estimation algorithm for the dispersion parameters the initial iterations will be traversing through regions of the parameter space that are relatively far from the final solution. These iterations do not need to have a precise estimate expected value. The later iterations will be close to the final solution, where a precise estimate of the expected value is required. This suggests that increasing the number of bootstrap samples from one iteration to the next could save computational effort and still deliver reasonable precision. A maximum number of bootstrap samples should be specified. This avoids spending too much effort in calculating the scores.

A simulation study was performed to assess the effect of choosing a varying number of bootstrap samples. A total of 10 data sets were generated and each data set was analysed 10 times using the bootstrap adjusted score method. Each data set consisted of $n = 100$ observations and $q = 35$ LASSO effects. The dispersion parameters were chosen to be $\phi = 0.1$ and $\sigma^2 = 1$. Four different numbers of bootstrap samples were used in each analysis: $B = 5$, $B = 25$, $B = 50$ and $B = 100$. Each of these were held constant throughout the estimation process. In addition to these four analyses, the increasing number of bootstrap samples scheme was employed with $B = 5$ initially and $B = \min(5 + 5i, 100)$ subsequently, where i is the iteration number.

The multiple estimates of each data set will exhibit variation due to the stochastic element introduced by the bootstrap. The variation for each bootstrap sample replication number for each data set was summarised by the standard deviation of the estimates (Figure 7.2). The CPU time taken to perform each analysis was also recorded (Figure 7.3). Not surprisingly, estimating using a larger number of bootstrap samples decreases the variability in the estimates. However, this comes at the cost of increased computation. Increasing the number of bootstrap samples after each iteration decreases the computational cost (compared with the large sample size) but retains the low estimation variance.

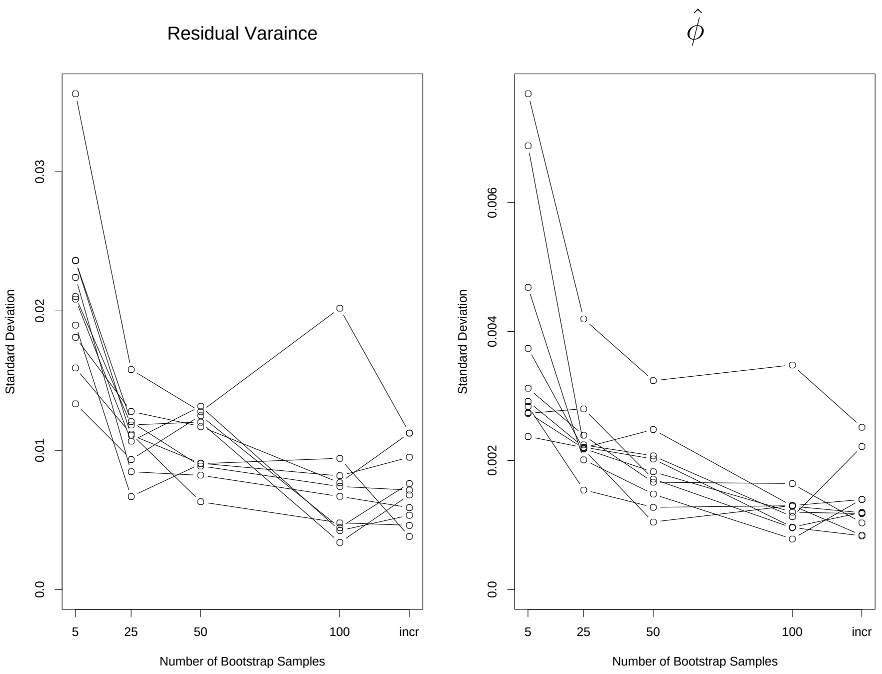

Figure 7.2: Standard deviation of the 10 estimates for each data set. The simulation values for dispersion parameters were $\sigma^2 = 1$ and $\phi = 0.1$. The label ‘incr’ represents the increasing bootstrap sample scheme.

These results indicate that increasing the number of bootstrap samples after each iteration decreases the computational effort required when compared to keeping the maximal number throughout the estimation process. However, for any particular data set, the incremental method does not seem to produce less variable results than the corresponding fixed sample number scheme (Figure 7.2). The estimates for the incremental scheme for some data sets have a slightly inflated standard deviation. Over all data sets the variability of the estimates does not seem to increase. This coupled with the savings in computation suggest that the incremental scheme should be used routinely.

7.7 Significance Testing of Estimates

After estimation of the dispersion parameters and the LASSO effects has been performed the task of the analyst is to interpret the model. The LASSO effects that have been estimated to be zero are easily interpreted - they do not explain variation in the observed outcomes. However, the non-zero estimates may be non-zero due to chance. Most commonly, a significance test is used to determine which effects are important and which effects are not. A testing strategy for the LASSO estimates is now described.

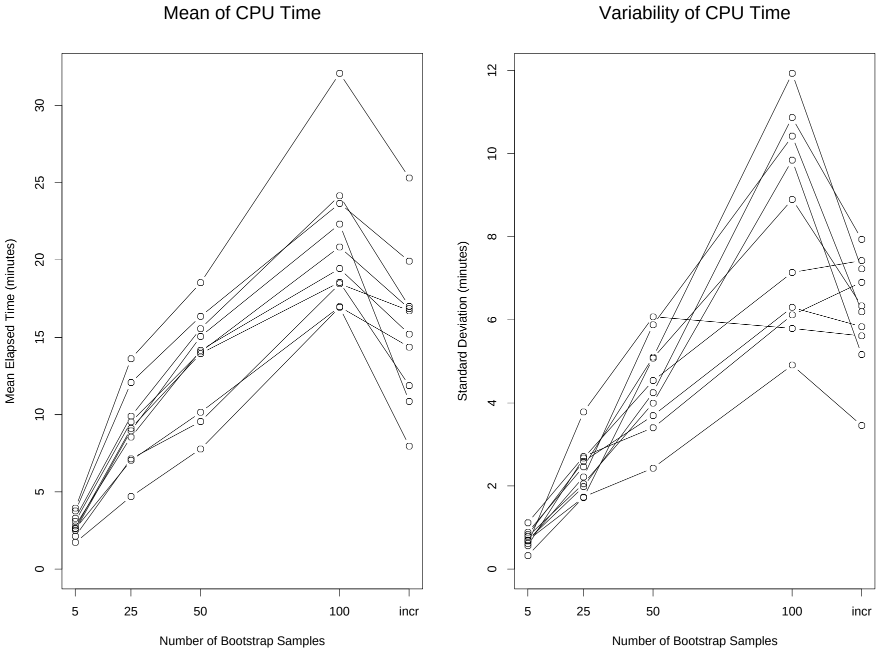

Figure 7.3: Mean (left panel) and standard deviation (right panel) time for estimation of dispersion parameters. The label 'incr' represents the increasing bootstrap sample scheme.

Standard errors of the LASSO estimates can be obtained using the approximations in Chapter 4. However, these are potentially misleading. They are based on a ridge regression (normal random effects model) approximation and may not be accurate. Also, the distribution of the LASSO estimates is unknown. Hence, the usefulness of standard errors is debatable. However, some asymptotic distribution results are known (Knight & Fu, 2000). These authors show that there is a positive probability mass at 0 in the asymptotic distribution if the true effect is zero (or small for small samples). Away from 0 the density is continuous. Unfortunately, these results require the unknown true LASSO effects. Knight & Fu (2000) present a bootstrap method for calculating the estimates' standard errors.

In this chapter, calculation of the distribution of the LASSO estimates is avoided and the LASSO estimates are tested directly using a simulation test (e.g. Davison & Hinkley, 1997, Chapter 4.2). This procedure generates a test of the null hypothesis that each LASSO effect is equal to zero. An empirical null distribution is generated by simulating data sets assuming that all the LASSO effects are zero and then analysing each simulated data set. For the model in (7.1.1) this requires simulation from a $\mathrm{N}_n(\mathbf{0}, \sigma^2\boldsymbol{I}_n)$ distribution. The value of σ^2 is unknown and is replaced by the estimate from the data $\hat{\sigma}^2$, obtained from the original analysis. The estimated dispersion parameters from the original analysis are used for predicting the LASSO effects for each simulated data set. Denote the estimates from the k^{th} simulated data set as $\bar{\boldsymbol{\beta}}^{(k)}$. The distribution of the simulated estimates $\bar{\boldsymbol{\beta}}^{(k)}$ forms the empirical null distribution.

Two types of (two-tailed) tests are available using this empirical null distribution. The first is a comparison-wise test where each effect is tested independently at the specified type-1 error rate α. A critical value for a single LASSO estimate $\bar{\beta}_i$ is found by obtaining the $100(1-\alpha)$ percentile of the empirical distribution of $|\bar{\beta}_i^{(k)}|$. If $|\bar{\beta}_i|$ is greater than this critical value then the effect is significant at the α level. This test does not take into account multiple testing; an especially important issue when there are a large number of LASSO effects and the probability of finding false positives is unacceptably high. The second type of test allows for multiple comparisons giving an experiment-wise test were a single critical value is used for all the LASSO estimates. The critical value is the $100(1-\alpha)$ percentile of the distribution of $\max(|\bar{\beta}_i^{(k)}|)$ where the maximum is taken of the q estimates for the k^{th} data set.

This procedure is illustrated for the prostate data used in Tibshirani (1996). This data is briefly explained and analysed in more detail in Section 7.9. Here, graphical representations of the empirical null distributions are given in Figures 7.4 and 7.5 for comparison-wise and experiment-wise tests respectively. The distributions are generated using 5000 simulations. This may be more than necessary for practical use, but it is beneficial for explanation. Perhaps the most striking feature of the distributions in Figure 7.4 is the spike at zero. This is not unexpected since the estimates $\bar{\boldsymbol{\beta}}^{(k)}$ are LASSO estimates and hence can have elements estimated to be identically zero.

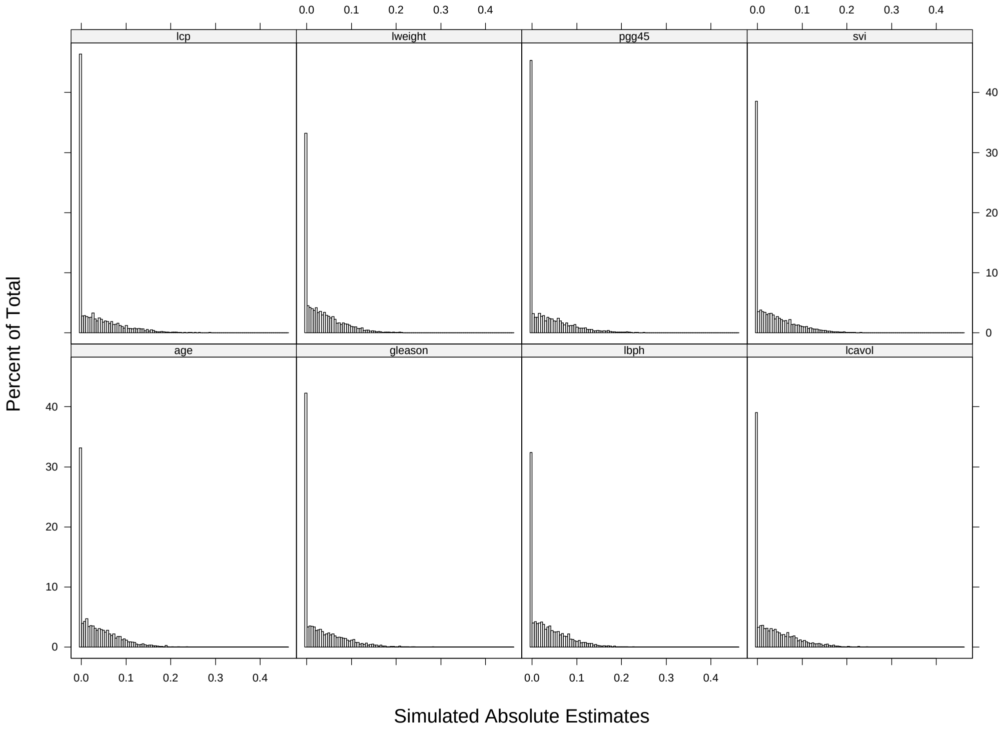

Figure 7.4: Empirical null distributions for the 8 comparison-wise tests for the prostate data.

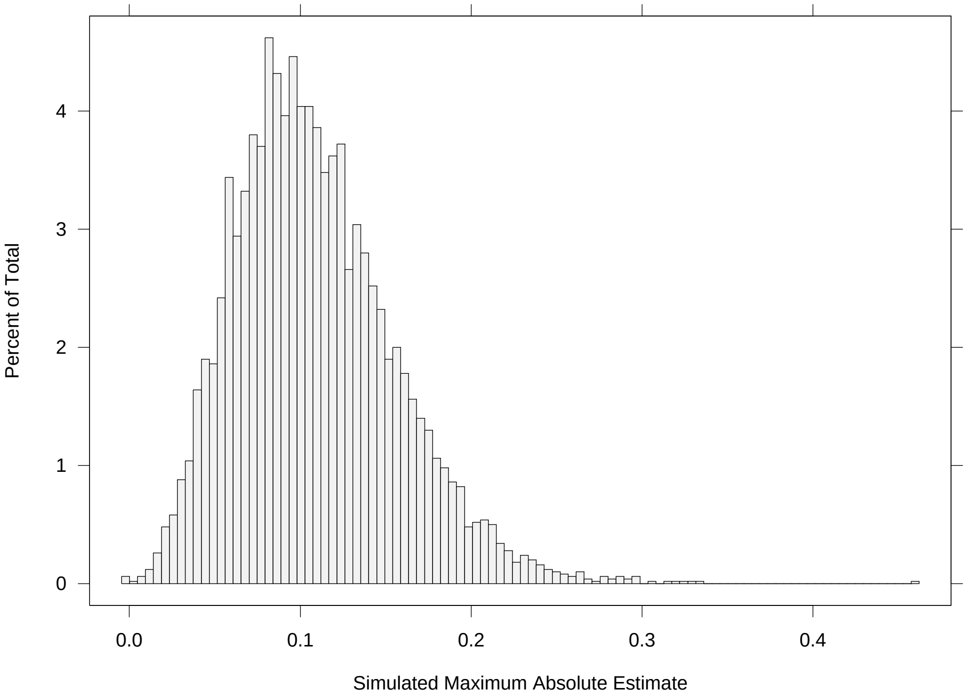

Figure 7.5: Empirical null distribution for the experiment-wise test for the prostate data. Critical value for a 5% experiment-wise test is 0.192.

7.8 Simulation Study

The performance of the LASSO random effects model with dispersion parameter estimation is assessed via simulation. Other types of LASSO model selection and other common subset selection methods are used as a comparison. In particular the random effects model LASSO is compared with estimates obtained from the adjusted score equations from Theorem 7.5 and from replacing q with $q_{\mathcal{S}}$ in Theorem 7.4 are considered. These are labeled LASSO$_{boot}$ and LASSO$_{qs}$ respectively. Other LASSO models, estimated by GCV$_{Tib}$ and GCV$_{fu}$ (see Chapter 6) and forwards selection, backwards selection, best subsets and full least squares are also considered (see Chapter 4). This is by no means an exhaustive list but it does provide the basis for a useful comparison to established methods.

The GCV statistics were evaluated at 200 different values of the constraint parameter t. These points were evenly spaced between 0 and $||\hat{\boldsymbol{\beta}}||_1$ where $\hat{\boldsymbol{\beta}}$ is the least squares estimate. The value of t that gives the minimum GCV statistic is used as the constraint for the final model.

Models using forwards selection, backwards selection and best subsets were chosen to minimise the C_p statistic. This is approximately equivalent to minimising Akaike's information criterion (Venables & Ripley, 2002, page 174).

7.8.1 *Simulation Design*

The random effects model in (7.1.1) was used in each of two related simulation studies. The simulation design was: number of observations $n = 100$, number of LASSO effects $q = 11$, the vector of LASSO effects $\boldsymbol{\beta}^T = (0, 0, 2, 0, 0, -5, 0, 0, 1, 0, 0)$ and the residuals were sampled independently from a $\mathrm{N}(0, 75)$ distribution. Each simulation study consisted of 100 simulated data sets. In the first simulation the rows of $\boldsymbol{L}$ were sampled from a $\mathrm{N}_{11}(\mathbf{0}, \boldsymbol{I})$ distribution. In the second simulation the rows of $\boldsymbol{L}$ were sampled from the normal distribution $\mathrm{N}_{11}(\mathbf{0}, \boldsymbol{A})$ where $\boldsymbol{A}$ was a correlation matrix corresponding to the structure obtained from an AR(1) process with correlation 0.819. The second simulation was designed to provide moderate ill-conditioning. The design matrices were held constant throughout both simulation studies. The conditioning number of the design matrices, the ratio of the largest to smallest eigenvalues, was 2.97 for the independent simulation and 105.61 for the correlated simulation.

The mean squared error for estimates was recorded (MSE= $(\boldsymbol{\beta} - \bar{\boldsymbol{\beta}})^T(\boldsymbol{\beta} - \bar{\boldsymbol{\beta}})/q$) for each simulated data set and its corresponding estimated model. Also recorded were the number of non-zero estimates, the number of significant effects (comparison-wise 5% level), the number of times the truly non-zero effects were significant in the model (true positives), the average number of other significant estimates (average false positives) and the estimates of the dispersion parameters. The estimate for the random effects dispersion was available only in the LASSO random effects models.

7.8.2 *Simulation Results*

Summaries of the simulations are given in Table 7.5. The LASSO methods, apart from GCV_{Tib}, produce models with lower MSE than all other methods. This was observed in both the independent and correlated simulations. The number of non-zero estimates was fewer for the LASSO_{boot} method than the other LASSO methods in both simulations. The LASSO_{boot} method seems to be finding the more important explanatory variables and including less of the others.

The residual variance estimate $\hat{\sigma}^2$ was higher in the correlated simulation than in the independent simulation for all methods. Since the total amount of variance in the outcomes must remain constant on average, the variance due to the explanatory variables must drop. This implies that the estimate of the dispersion parameter $(\hat{\phi})$ should decrease between the independent and correlated simulations. This occurs for the LASSO_{boot} method but does not for the LASSO_{qs} method. This attribute of the LASSO_{qs} method is a concern and potentially reflects its ad-hoc motivation.

The number of true and false positives are presented in Table 7.6. In the independent simulation, all methods excluding backwards selection, had approximately equal power to detect all three explanatory variables affecting the outcome. The rate of false positives approximately matches the nominal level for all methods. In the correlated simulation the least squares method had poor power to detect the most important explanatory variable (l_6). The backwards selection and best subsets methods exhibited good power but also suffered from a highly inflated false positive rate. Both the LASSO methods had reasonable ability to identify the important explanatory variables and suffered only slightly from an increase in the rate of false positives.

Table 7.5: Summary of simulation studies with independent and correlated $\boldsymbol{L}$. MSE$_e$ is the average mean squared error for estimates, $q_{\mathcal{S}}$ and $q_{\mathcal{S}}^*$ are the average number of non-zero and significant estimates respectively, and $\hat{\sigma}^2$ and $\hat{\phi}$ are mean values for the dispersion parameter estimates.

		MSE	$q_{\mathcal{S}}$	$q_{\mathcal{S}}^*$	$\hat{\sigma}^2$	$\hat{\phi}$
Independent $\boldsymbol{L}$	LASSO GCV$_{Tib}$	0.56 (0.03)	9.95 (0.09)	-	-	-
	LASSO GCV$_{Fu}$	0.43 (0.02)	6.22 (0.14)	-	-	-
	LASSO$_{qs}$	0.45 (0.02)	6.42 (0.13)	2.18 (0.08)	71.44 (1.0)	1.31 (0.03)
	LASSO$_{boot}$	0.44 (0.02)	5.64 (0.20)	2.22 (0.08)	71.51 (1.0)	1.05 (0.02)
	Forward Select	0.55 (0.03)	2.68 (0.10)	2.23 (0.09)	70.89 (1.0)	-
	Backward Select	0.64 (0.03)	3.60 (0.13)	2.39 (0.12)	69.87 (1.0)	-
	Best Subsets	0.64 (0.03)	3.55 (0.12)	2.32 (0.09)	69.90 (1.0)	-
	Least Squares	0.84 (0.03)	11.00 (0.00)	2.16 (0.08)	72.58 (1.1)	-
Correlated $\boldsymbol{L}$	LASSO GCV$_{Tib}$	2.46 (0.21)	9.56 (0.10)	-	-	-
	LASSO GCV$_{Fu}$	1.52 (0.14)	5.51 (0.15)	-	-	-
	LASSO$_{qs}$	1.36 (0.10)	3.67 (0.17)	1.82 (0.11)	76.74 (1.2)	1.34 (0.06)
	LASSO$_{boot}$	1.44 (0.09)	2.80 (0.19)	1.89 (0.12)	77.15 (1.2)	0.90 (0.04)
	Forward Select	2.49 (0.23)	2.62 (0.10)	2.17 (0.08)	73.89 (1.1)	-
	Backward Select	3.68 (0.32)	3.41 (0.14)	2.62 (0.13)	72.97 (1.1)	-
	Best Subsets	3.43 (0.32)	3.24 (0.14)	2.63 (0.12)	73.02 (1.1)	-
	Least Squares	4.93 (0.36)	11.00 (0.00)	1.32 (0.11)	76.26 (1.2)	-

Parenthetic values are standard errors over simulations

Table 7.6: Number of true and false positives from simulation studies (100 simulations each). The true positives correspond to the number of times the 3rd, 6th and 9th explanatory variables (l_3, l_6 and l_9) were significant. The false positives (Other) are the average (over variables) of all the other significant effects. The comparison wise error rate was set at 5%.

	Independent L				Correlated L			
	l_3	l_6	l_9	Other	l_3	l_6	l_9	Other
LASSO$_{qs}$	66	100	13	4.9	23	82	13	8.0
LASSO$_{boot}$	67	100	14	5.1	21	83	14	8.9
Forward Select	66	100	14	5.4	33	82	16	10.8
Backward Select	66	90	23	7.5	37	78	20	15.9
Best Subsets	66	100	18	6.0	39	82	21	15.1
Least Squares	66	100	15	4.4	15	50	15	6.5

In summary, if only MSE is considered, then GCV$_{Fu}$, LASSO$_{qs}$ or LASSO$_{boot}$ are superior to GCV$_{Tib}$, full least squares and the selection methods. These three methods had approximately equal rates of true and false positives. However, the LASSO$_{boot}$ method had the lowest number of non-zero estimates. For this reason it should be preferred.

7.9 Example: Prostate Data

The prostate data was used as an example by Tibshirani (1996), Fu (1998) and Osborne et al. (2000) and is re-analysed in this section. These data come from a study by Stamey et al. (1989) who examined the relationship between the level of prostate specific antigen and eight clinical measures on men who were about to receive a radical prostatectomy. The clinical measures used as explanatory variables were log(cancer volume) (lcavol), log(prostate weight) (lweight), age, log(bengin prostatic hyperplasia amount) (lbph), seminal vesicle invasion (svi), log(capsular penetration) (lcp), Gleason score (gleason) and percentage Gleason scores 4 or (pgg45). For more details about the study see Stamey et al. (1989). The clinical measures and the outcome are centred and the clinical measures are also standardised. The model fitted was a main effects only model. It has been assumed that the explanatory variables have been transformed appropriately. This assumption enables direct comparison to previous analyses.

Osborne et al. (2000, Remark 10) pointed out that there was an error in the analysis by Tibshirani (1996). This error had the effect of choosing a model by GCV$_{Tib}$ that had a too severe constraint. The results obtained from the analysis presented here (Table 7.7) agree closely with those from Osborne et al. (2000).

The model was fitted to the data using all the methods described in the simulation study. The same model was identified using backwards selection and best subsets so results are only presented for backwards selection. Significance tests for all methods were performed using comparison-wise tests. For the LASSO random effects model estimation methods, 5000 data sets were simulated to produce the empirical null distributions. No attempt was made to assess the significance of the terms from the GCV models as no obvious estimate of residual variance was available.

Table 7.7: Estimates for the different estimation methods for the prostate cancer data example.

	LASSO				Selection		
	$\text{GCV}^{\ddagger}_{Tib}$	$\text{GCV}^{\ddagger}_{Fu}$	LASSO_{qs}	LASSO_{boot}	Forwards	Backwards	Full Least Squares
lcavol	0.629	0.621	0.626***	0.627***	0.650***	0.648***	0.692***
lweight	0.205	0.190	0.200***	0.201**	0.253***	0.194*	0.226**
age	-0.083	-0.047	-0.070	-0.073	-	-	$-0.146^{\dagger}$
lbph	0.124	0.103	$0.116^{\dagger}$	$0.118^{\dagger}$	-	0.131	$0.155^{\dagger}$
svi	0.256	0.246	0.252**	0.253***	0.276**	0.295***	0.317**
lcp	-0.002	0.000	0.000	0.000	-	-	-0.147
gleason	0.009	0.000	0.004	0.006	-	-	0.033
pgg45	0.073	0.063	0.070	0.071	-	-	0.128
σ^2	-	-	0.504	0.504	0.514	0.505	0.502

‡ Significance tests not available.

Comparison-wise significance $^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$ and $^{\dagger}p < 0.1$.

The four different methods for LASSO estimation, namely the two GCV methods and the two random effects model methods, behaved similarly. Only small differences in the size of the effects were seen although they did have differing numbers of zero effects. Qualitatively this appears important while quantitatively it is not. The significance levels for the two random LASSO models did not differ substantially.

For these data the behaviour of the LASSO method lay between the selection methods and the full least squares method. The LASSO identified four significant terms (lcavol, lweight, lbph and svi). The subset methods identified only the three significant terms while backwards selection and best subsets included lbph in the model even though it was not significant.

7.10 Summary

In this chapter, a method of estimation for the dispersion parameters in the LASSO model was presented. Estimation was based on the score equations from an approximate likelihood. These score equations were biased and were adjusted using an empirical estimate of the scores' expected values. Inference was performed by using a simulation method, which empirically generates a testing distribution for the LASSO effects. The methodology presented in this chapter was demonstrated by simulation to be competitive against other LASSO methods and subset selection methods.

Chapter 8

The LASSO Linear Mixed Model

8.1 Introduction

Previously the LASSO model was specified as a random effects model (Chapters 5 and 7). Under this formulation the parameters to be estimated were dispersion parameters and the LASSO estimates were predictions from a statistical model. This method of estimation was shown via simulation to be a competitive and useful method for estimating the LASSO effects. However, the random effects model is too restrictive for many real applications where models allowing for fixed effects and random normal effects are required.

Accommodation of fixed, random normal and random LASSO effects is facilitated by embedding the LASSO effects into a standard mixed model to produce the LASSO linear mixed model (LLMM).

$$\boldsymbol{y} = \boldsymbol{X\tau} + \boldsymbol{Zu} + \boldsymbol{L\beta} + \boldsymbol{e} \tag{8.1.1}$$

where

$$\begin{aligned} \boldsymbol{u} &\sim \mathrm{N}(\boldsymbol{0}, \sigma^2 \boldsymbol{G}(\boldsymbol{\gamma})), \\ \beta_i &\sim \mathrm{DE}(0, 2\phi^2) \qquad i = 1, 2, \ldots, q, \\ \boldsymbol{e} &\sim \mathrm{N}(\boldsymbol{0}, \sigma^2 \boldsymbol{R}(\boldsymbol{\theta})), \end{aligned}$$

σ^2 is the residual variance, $\boldsymbol{G}(\boldsymbol{\gamma})$ is the scaled variance matrix of the random normal effects, $2\phi^2$ is the variance of the LASSO effects and $\boldsymbol{R}(\boldsymbol{\theta})$ is the scaled variance matrix for the residuals. The vectors $\boldsymbol{\gamma}$ and $\boldsymbol{\theta}$ completely parameterise the correlation matrices $\boldsymbol{G}$ and $\boldsymbol{R}$ respectively. Often the matrix $\boldsymbol{R}$ is assumed to be the identity, an assumption not made in this chapter. However, the S-PLUS code written to run the models presented in this chapter does not allow estimation of these parameters. This code is presented in Appendix D.

The topic of this chapter is the estimation of parameters $\boldsymbol{\tau}$, $\boldsymbol{\gamma}$, ϕ and $\boldsymbol{\theta}$ and the prediction of the random effects $\boldsymbol{u}$ and $\boldsymbol{\beta}$. The model effects $\boldsymbol{\tau}$, $\boldsymbol{u}$ and $\boldsymbol{\beta}$ are estimated/predicted using an extension of the mixed model equations (Henderson, 1950). These equations are solved using a modified Gaussian elimination technique. The modification is necessary as solving for $\boldsymbol{\beta}$ requires non-standard methodology (Section 6.3). The dispersion parameters are estimated using an approximation to the restricted likelihood. The approximation is obtained using the partial Laplace method described in Chapter 3. Special consideration is required in this context as the exponent of the joint distribution is not differentiable and the normal approximation may be singular. These are overcome using a similar strategy to that outlined in Chapter 7. For comparison, the standard Laplace approximation to the unrestricted likelihood is also presented. Differences similar to those between the restricted likelihood and the un-restricted likelihood for standard models are observed.

Solving the score equations requires iteration as the scores are functions of the LASSO estimates $\bar{\boldsymbol{\beta}}$. A further complication arises at each iteration as there is no closed form for the estimates of $\boldsymbol{\gamma}$ and $\boldsymbol{\theta}$. These dispersion parameters are updated using a scoring algorithm based on an average information algorithm (Chapter 3), which is more efficient than updating using a standard Newton-Raphson algorithm.

8.2 The LASSO Mixed Model Equations

The location effects in the LLMM are estimated by maximising the so-called joint distribution of the observed outcomes and the unobserved random LASSO and random normal effects. The dispersion parameters are assumed known or are replaced by appropriate plug-in estimators. This parallels the estimation of fixed and random effects in the standard mixed model (Henderson, 1950). In the standard mixed model the mode of the predictive distribution (posterior in Bayesian terminology) coincides with the mean, and hence finding the maximum (mode) of the joint distribution also finds the mean. This is not the case for the LASSO linear mixed model and so the estimates are not best in the sense that they are expected values. The ability of the LASSO to estimate effects to be identically zero is precisely due to the estimates being modes (see Chapter 5).

The joint distribution of the outcomes, the random normal effects and the random LASSO effects is

$$\begin{aligned}
f(\boldsymbol{y},\boldsymbol{u},\boldsymbol{\beta}) &= f(\boldsymbol{y}|\boldsymbol{u},\boldsymbol{\beta})f(\boldsymbol{u})f(\boldsymbol{\beta}) \\
&= \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}}|\boldsymbol{R}|^{-\frac{1}{2}}\exp\left(-\frac{1}{2\sigma^2}(\boldsymbol{y}-\boldsymbol{X\tau}-\boldsymbol{Zu}-\boldsymbol{L\beta})^T\boldsymbol{R}^{-1}(\boldsymbol{y}-\boldsymbol{X\tau}-\boldsymbol{Zu}-\boldsymbol{L\beta})\right) \\
&\qquad\qquad \times\frac{1}{(2\pi\sigma^2)^{\frac{r}{2}}}|\boldsymbol{G}|^{-\frac{1}{2}}\exp\left(-\frac{1}{2\sigma^2}\boldsymbol{u}^T\boldsymbol{G}^{-1}\boldsymbol{u}\right)\frac{1}{(2\phi)^q}\exp\left(-\frac{1}{\phi}\|\boldsymbol{\beta}\|_1\right).
\end{aligned}$$

The estimating equations for $\boldsymbol{\tau}$ and $\boldsymbol{u}$ are derived by taking the derivatives of the logarithm of the joint distribution and assigning them to zero. The joint distribution is not differentiable w.r.t. $\boldsymbol{\beta}$. However, the joint distribution is convex in $\boldsymbol{\beta}$ and, as in Chapter 7, the sub-differential is used. The sub-differential is equated to zero to obtain the estimating equation for $\boldsymbol{\beta}$. The derivatives w.r.t. $\boldsymbol{\tau}$ and $\boldsymbol{u}$, the sub-differential w.r.t. $\boldsymbol{\beta}$ are

$$\begin{aligned}
\frac{\partial f(\boldsymbol{y},\boldsymbol{u},\boldsymbol{\beta})}{\partial \boldsymbol{\tau}} &= \frac{1}{\sigma^2}\boldsymbol{X}^T\boldsymbol{R}^{-1}\boldsymbol{y} - \frac{1}{\sigma^2}\left(\boldsymbol{X}^T\boldsymbol{R}^{-1}\boldsymbol{X}\boldsymbol{\tau} + \boldsymbol{X}^T\boldsymbol{R}^{-1}\boldsymbol{Z}\boldsymbol{u} + \boldsymbol{X}^T\boldsymbol{R}^{-1}\boldsymbol{L}\boldsymbol{\beta}\right)\\
\frac{\partial f(\boldsymbol{y},\boldsymbol{u},\boldsymbol{\beta})}{\partial \boldsymbol{u}} &= \frac{1}{\sigma^2}\boldsymbol{Z}^T\boldsymbol{R}^{-1}\boldsymbol{y} - \frac{1}{\sigma^2}\left(\boldsymbol{Z}^T\boldsymbol{R}^{-1}\boldsymbol{X}\boldsymbol{\tau} + (\boldsymbol{Z}^T\boldsymbol{R}^{-1}\boldsymbol{Z} + \boldsymbol{G}^{-1})\boldsymbol{u} + \boldsymbol{Z}^T\boldsymbol{R}^{-1}\boldsymbol{L}\boldsymbol{\beta}\right)\\
\partial_{\boldsymbol{\beta}} f(\boldsymbol{y},\boldsymbol{u},\boldsymbol{\beta}) &= \frac{1}{\sigma^2}\boldsymbol{L}^T\boldsymbol{R}^{-1}\boldsymbol{y} - \frac{1}{\phi}\boldsymbol{v} - \frac{1}{\sigma^2}\left(\boldsymbol{L}^T\boldsymbol{R}^{-1}\boldsymbol{X}\boldsymbol{\tau} + \boldsymbol{L}^T\boldsymbol{R}^{-1}\boldsymbol{Z}\boldsymbol{u} + \boldsymbol{L}^T\boldsymbol{R}^{-1}\boldsymbol{L}\boldsymbol{\beta}\right)
\end{aligned}$$

and the estimating equations are

$$\begin{aligned}
\boldsymbol{X}^T\boldsymbol{R}^{-1}\boldsymbol{X}\hat{\boldsymbol{\tau}} + \boldsymbol{X}^T\boldsymbol{R}^{-1}\boldsymbol{Z}\tilde{\boldsymbol{u}} + \boldsymbol{X}^T\boldsymbol{R}^{-1}\boldsymbol{L}\bar{\boldsymbol{\beta}} &= \boldsymbol{X}^T\boldsymbol{R}^{-1}\boldsymbol{y}\\
\boldsymbol{Z}^T\boldsymbol{R}^{-1}\boldsymbol{X}\hat{\boldsymbol{\tau}} + (\boldsymbol{Z}^T\boldsymbol{R}^{-1}\boldsymbol{Z} + \boldsymbol{G}^{-1})\tilde{\boldsymbol{u}} + \boldsymbol{Z}^T\boldsymbol{R}^{-1}\boldsymbol{L}\bar{\boldsymbol{\beta}} &= \boldsymbol{Z}^T\boldsymbol{R}^{-1}\boldsymbol{y}\\
\boldsymbol{L}^T\boldsymbol{R}^{-1}\boldsymbol{X}\hat{\boldsymbol{\tau}} + \boldsymbol{L}^T\boldsymbol{R}^{-1}\boldsymbol{Z}\tilde{\boldsymbol{u}} + \boldsymbol{L}^T\boldsymbol{R}^{-1}\boldsymbol{L}\bar{\boldsymbol{\beta}} &= \boldsymbol{L}^T\boldsymbol{R}^{-1}\boldsymbol{y} - \frac{\sigma^2}{\phi}\bar{\boldsymbol{v}}.
\end{aligned}$$

These estimating equations are called the LASSO mixed model equations (LMME). Simultaneous estimates of $\boldsymbol{\tau}$, $\boldsymbol{u}$ and $\boldsymbol{\beta}$ are obtained by replacing the unknown effects with their estimates and then solving the resulting simultaneous equations. The LMME can be expressed in matrix notation

$$\begin{pmatrix}
\boldsymbol{L}^T\boldsymbol{R}^{-1}\boldsymbol{L} & \boldsymbol{L}^T\boldsymbol{R}^{-1}\boldsymbol{X} & \boldsymbol{L}^T\boldsymbol{R}^{-1}\boldsymbol{Z}\\
\boldsymbol{X}^T\boldsymbol{R}^{-1}\boldsymbol{L} & \boldsymbol{X}^T\boldsymbol{R}^{-1}\boldsymbol{X} & \boldsymbol{X}^T\boldsymbol{R}^{-1}\boldsymbol{Z}\\
\boldsymbol{Z}^T\boldsymbol{R}^{-1}\boldsymbol{L} & \boldsymbol{Z}^T\boldsymbol{R}^{-1}\boldsymbol{X} & \boldsymbol{Z}^T\boldsymbol{R}^{-1}\boldsymbol{Z} + \boldsymbol{G}^{-1}
\end{pmatrix}
\begin{pmatrix}\bar{\boldsymbol{\beta}}\\ \hat{\boldsymbol{\tau}}\\ \tilde{\boldsymbol{u}}\end{pmatrix}
=
\begin{pmatrix}
\boldsymbol{L}^T\boldsymbol{R}^{-1}\boldsymbol{y} - \frac{\sigma^2}{\phi}\bar{\boldsymbol{v}}\\
\boldsymbol{X}^T\boldsymbol{R}^{-1}\boldsymbol{y}\\
\boldsymbol{Z}^T\boldsymbol{R}^{-1}\boldsymbol{y}
\end{pmatrix}. \tag{8.2.1}$$

The LMME are solved using a modified version of Gaussian elimination. The modification is needed because the interior point algorithm for correlated effects in Chapter 6 has to be used to estimate $\bar{\boldsymbol{\beta}}$ and $\bar{\boldsymbol{v}}$. During the absorption process in Gaussian elimination, the residual sums of squares can be obtained if the coefficient matrix on the left hand side of (8.2.1) is appended with the outcome vector on the right hand side. This leads to the augmented LMME matrix

$$\begin{pmatrix}
\boldsymbol{y}^T\boldsymbol{R}^{-1}\boldsymbol{y} & \boldsymbol{y}^T\boldsymbol{R}^{-1}\boldsymbol{L} - \frac{\sigma^2}{\phi}\bar{\boldsymbol{v}}^T & \boldsymbol{y}^T\boldsymbol{R}^{-1}\boldsymbol{X} & \boldsymbol{y}^T\boldsymbol{R}^{-1}\boldsymbol{Z}\\
\boldsymbol{L}^T\boldsymbol{R}^{-1}\boldsymbol{y} - \frac{\sigma^2}{\phi}\bar{\boldsymbol{v}} & \boldsymbol{L}^T\boldsymbol{R}^{-1}\boldsymbol{L} & \boldsymbol{L}^T\boldsymbol{R}^{-1}\boldsymbol{X} & \boldsymbol{L}^T\boldsymbol{R}^{-1}\boldsymbol{Z}\\
\boldsymbol{X}^T\boldsymbol{R}^{-1}\boldsymbol{y} & \boldsymbol{X}^T\boldsymbol{R}^{-1}\boldsymbol{L} & \boldsymbol{X}^T\boldsymbol{R}^{-1}\boldsymbol{X} & \boldsymbol{X}^T\boldsymbol{R}^{-1}\boldsymbol{Z}\\
\boldsymbol{Z}^T\boldsymbol{R}^{-1}\boldsymbol{y} & \boldsymbol{Z}^T\boldsymbol{R}^{-1}\boldsymbol{L} & \boldsymbol{Z}^T\boldsymbol{R}^{-1}\boldsymbol{X} & \boldsymbol{Z}^T\boldsymbol{R}^{-1}\boldsymbol{Z} + \boldsymbol{G}^{-1}
\end{pmatrix}.$$

The Gaussian elimination process starts with the absorption of the last row of the LMME. For example the term $\boldsymbol{X}^T\boldsymbol{R}^{-1}\boldsymbol{X}$ becomes

$$\begin{aligned}
\boldsymbol{X}^T\boldsymbol{R}^{-1}\boldsymbol{X} \leftarrow \boldsymbol{X}^T\boldsymbol{R}^{-1}\boldsymbol{X} &- \boldsymbol{X}^T\boldsymbol{R}^{-1}\boldsymbol{Z}(\boldsymbol{Z}^T\boldsymbol{R}^{-1}\boldsymbol{Z} + \boldsymbol{G}^{-1})^{-1}\boldsymbol{Z}^T\boldsymbol{R}^{-1}\boldsymbol{X}\\
&= \boldsymbol{X}^T\left(\boldsymbol{R}^{-1} - \boldsymbol{R}^{-1}\boldsymbol{Z}(\boldsymbol{Z}^T\boldsymbol{R}^{-1}\boldsymbol{Z} + \boldsymbol{G}^{-1})^{-1}\boldsymbol{Z}^T\boldsymbol{R}^{-1}\right)\boldsymbol{X}\\
&= \boldsymbol{X}^T\boldsymbol{H}^{-1}\boldsymbol{X}
\end{aligned}$$

where $\boldsymbol{H} = \boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}^T + \boldsymbol{R}$. The updated system of equations is

$$\begin{pmatrix} \boldsymbol{y}^T\boldsymbol{H}^{-1}\boldsymbol{y} & \boldsymbol{y}^T\boldsymbol{H}^{-1}\boldsymbol{L} - \frac{\sigma^2}{\phi}\bar{\boldsymbol{v}}^T & \boldsymbol{y}^T\boldsymbol{H}^{-1}\boldsymbol{X} & \boldsymbol{0} \\ \boldsymbol{L}^T\boldsymbol{H}^{-1}\boldsymbol{y} - \frac{\sigma^2}{\phi}\bar{\boldsymbol{v}} & \boldsymbol{L}^T\boldsymbol{H}^{-1}\boldsymbol{L} & \boldsymbol{L}^T\boldsymbol{H}^{-1}\boldsymbol{X} & \boldsymbol{0} \\ \boldsymbol{X}^T\boldsymbol{H}^{-1}\boldsymbol{y} & \boldsymbol{X}^T\boldsymbol{H}^{-1}\boldsymbol{L} & \boldsymbol{X}^T\boldsymbol{H}^{-1}\boldsymbol{X} & \boldsymbol{0} \\ \boldsymbol{Z}^T\boldsymbol{R}^{-1}\boldsymbol{y} & \boldsymbol{Z}^T\boldsymbol{R}^{-1}\boldsymbol{L} & \boldsymbol{Z}^T\boldsymbol{R}^{-1}\boldsymbol{X} & \boldsymbol{Z}^T\boldsymbol{R}^{-1}\boldsymbol{Z} + \boldsymbol{G}^{-1} \end{pmatrix}.$$

The absorption process is performed again, this time on the second last row. For example the term $\boldsymbol{L}^T\boldsymbol{H}^{-1}\boldsymbol{L}$ becomes

$$\begin{aligned} \boldsymbol{L}^T\boldsymbol{H}^{-1}\boldsymbol{L} &\leftarrow \boldsymbol{L}^T\boldsymbol{H}^{-1}\boldsymbol{L} - \boldsymbol{L}^T\boldsymbol{H}^{-1}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{H}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{L} \\ &= \boldsymbol{L}^T\boldsymbol{P}\boldsymbol{L} \end{aligned}$$

where $\boldsymbol{P} = \boldsymbol{H}^{-1} - \boldsymbol{H}^{-1}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{H}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{H}^{-1}$. The updated system of equations is

$$\begin{pmatrix} \boldsymbol{y}^T\boldsymbol{P}\boldsymbol{y} & \boldsymbol{y}^T\boldsymbol{P}\boldsymbol{L} - \frac{\sigma^2}{\phi}\bar{\boldsymbol{v}}^T & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{L}^T\boldsymbol{P}\boldsymbol{y} - \frac{\sigma^2}{\phi}\bar{\boldsymbol{v}} & \boldsymbol{L}^T\boldsymbol{P}\boldsymbol{L} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{X}^T\boldsymbol{H}^{-1}\boldsymbol{y} & \boldsymbol{X}^T\boldsymbol{H}^{-1}\boldsymbol{L} & \boldsymbol{X}^T\boldsymbol{H}^{-1}\boldsymbol{X} & \boldsymbol{0} \\ \boldsymbol{Z}^T\boldsymbol{R}^{-1}\boldsymbol{y} & \boldsymbol{Z}^T\boldsymbol{R}^{-1}\boldsymbol{L} & \boldsymbol{Z}^T\boldsymbol{R}^{-1}\boldsymbol{X} & \boldsymbol{Z}^T\boldsymbol{R}^{-1}\boldsymbol{Z} + \boldsymbol{G}^{-1} \end{pmatrix}. \tag{8.2.2}$$

The LASSO estimates are straight forward and are identical in form to those given in (6.2.2)

$$\boldsymbol{L}^T\boldsymbol{P}\boldsymbol{L}\bar{\boldsymbol{\beta}} = \boldsymbol{L}^T\boldsymbol{P}\boldsymbol{y} - \frac{\sigma^2}{\phi}\bar{\boldsymbol{v}}$$

giving

$$\bar{\boldsymbol{\beta}} = (\boldsymbol{L}^T\boldsymbol{P}\boldsymbol{L})^{-}\boldsymbol{L}^T\boldsymbol{P}\boldsymbol{y} - \frac{\sigma^2}{\phi}(\boldsymbol{L}^T\boldsymbol{P}\boldsymbol{L})^{-}\bar{\boldsymbol{v}}. \tag{8.2.3}$$

The last step of absorption gives the residual sums of squares as the first element (e.g. Seber, 1977 and Gilmour et al., 1995). The matrix $\boldsymbol{L}^T\boldsymbol{P}\boldsymbol{L}$ may be non-singular if $q > n - p$, but for ease of illustration assume that $\boldsymbol{L}^T\boldsymbol{P}\boldsymbol{L}$ is non-singular. Using the functional form for the LASSO estimates in (8.2.3)

$$\begin{aligned} \boldsymbol{y}^T\boldsymbol{P}\boldsymbol{y} &\leftarrow \boldsymbol{y}^T\boldsymbol{P}\boldsymbol{y} - \left(\boldsymbol{L}^T\boldsymbol{P}\boldsymbol{y} - \frac{\sigma^2}{\phi}\bar{\boldsymbol{v}}\right)^T (\boldsymbol{L}^T\boldsymbol{P}\boldsymbol{L})^{-1}\left(\boldsymbol{L}^T\boldsymbol{P}\boldsymbol{y} - \frac{\sigma^2}{\phi}\bar{\boldsymbol{v}}\right) \\ &= \boldsymbol{y}^T\boldsymbol{P}\boldsymbol{y} - \left(\left(\boldsymbol{L}^T\boldsymbol{P}\boldsymbol{y} - \frac{\sigma^2}{\phi}\bar{\boldsymbol{v}}\right)^T (\boldsymbol{L}^T\boldsymbol{P}\boldsymbol{L})^{-1}\right)\boldsymbol{L}^T\boldsymbol{P}\boldsymbol{L}\left((\boldsymbol{L}^T\boldsymbol{P}\boldsymbol{L})^{-1}\left(\boldsymbol{L}^T\boldsymbol{P}\boldsymbol{y} - \frac{\sigma^2}{\phi}\bar{\boldsymbol{v}}\right)\right) \\ &= \boldsymbol{y}^T\boldsymbol{P}\boldsymbol{y} - \bar{\boldsymbol{\beta}}^T\boldsymbol{L}\boldsymbol{P}\boldsymbol{L}\bar{\boldsymbol{\beta}}. \end{aligned} \tag{8.2.4}$$

The estimates $\hat{\boldsymbol{\tau}}$ and $\tilde{\boldsymbol{u}}$ can now be given as solutions to the absorbed matrix (8.2.2). The fixed and random normal effects' estimates are found by back-substitution. Consider the back substitution for the fixed effects estimates $\hat{\boldsymbol{\tau}}$

$$\boldsymbol{X}^T\boldsymbol{H}^{-1}\boldsymbol{L}\bar{\boldsymbol{\beta}} + \boldsymbol{X}^T\boldsymbol{H}^{-1}\boldsymbol{X}\hat{\boldsymbol{\tau}} = \boldsymbol{X}^T\boldsymbol{H}^{-1}\boldsymbol{y}$$

giving

$$\hat{\boldsymbol{\tau}} = (\boldsymbol{X}^T\boldsymbol{H}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{H}^{-1}(\boldsymbol{y} - \boldsymbol{L}\bar{\boldsymbol{\beta}}).$$

This is the generalised least squares estimate based on the outcomes adjusted for the LASSO estimates. For comparison to the standard mixed model see Theorem 3.1. The next step in back substitution gives the random effect predictions

$$\boldsymbol{Z}^T\boldsymbol{R}^{-1}\boldsymbol{L}\bar{\boldsymbol{\beta}}+\boldsymbol{Z}^T\boldsymbol{R}^{-1}\boldsymbol{X}\hat{\boldsymbol{\tau}} + (\boldsymbol{Z}^T\boldsymbol{R}^{-1}\boldsymbol{Z} + \boldsymbol{G}^{-1})\tilde{\boldsymbol{u}} = \boldsymbol{Z}^T\boldsymbol{R}^{-1}\boldsymbol{y}$$

giving

$$\begin{aligned}\tilde{\boldsymbol{u}} &= \left(\boldsymbol{Z}^T\boldsymbol{R}^{-1}\boldsymbol{Z} + \boldsymbol{G}^{-1}\right)^{-1}\boldsymbol{Z}^T\boldsymbol{R}^{-1}(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\tau}}\boldsymbol{L}\bar{\boldsymbol{\beta}})\\ &= \left(\boldsymbol{Z}^T\boldsymbol{R}^{-1}\boldsymbol{Z} + \boldsymbol{G}^{-1}\right)^{-1}\boldsymbol{Z}^T\boldsymbol{R}^{-1}\\ &\qquad\qquad \times \left(\boldsymbol{y} - \boldsymbol{L}\bar{\boldsymbol{\beta}} - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{H}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{H}^{-1}(\boldsymbol{y} - \boldsymbol{L}\bar{\boldsymbol{\beta}})\right)\\ &= \left(\boldsymbol{Z}^T\boldsymbol{R}^{-1}\boldsymbol{Z} + \boldsymbol{G}^{-1}\right)^{-1}\boldsymbol{Z}^T\boldsymbol{R}^{-1}\\ &\qquad\qquad \times \left(\boldsymbol{H}\left(\boldsymbol{H}^{-1} - \boldsymbol{H}^{-1}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{H}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{H}^{-1}\right)(\boldsymbol{y} - \boldsymbol{L}\bar{\boldsymbol{\beta}})\right)\end{aligned}$$

using Result B.2

$$\begin{aligned}&= \boldsymbol{G}\boldsymbol{Z}^T(\boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}^T + \boldsymbol{R})^{-1}\boldsymbol{H}\boldsymbol{P}(\boldsymbol{y} - \boldsymbol{L}\bar{\boldsymbol{\beta}})\\ &= \boldsymbol{G}\boldsymbol{Z}^T\boldsymbol{P}(\boldsymbol{y} - \boldsymbol{L}\bar{\boldsymbol{\beta}}).\end{aligned}$$

This is analogous to the random effect estimate for the standard mixed model from Theorem 3.1. However, like the fixed effect estimates, the outcomes are adjusted for the LASSO estimates.

8.3 Approximate Marginal and Restricted Marginal Likelihood

For a model containing only LASSO random effects (Chapter 7), the likelihood was based on a Laplace approximation of the marginal distribution. The same method can be applied here producing a marginal likelihood that assumes that the fixed effects are known. Results for the standard mixed model with only fixed and random normal effects show that such a likelihood produces biased estimates of the dispersion parameters. Hence, a restricted likelihood (Patterson & Thompson, 1971) is often employed. The restriction is a linear transformation of the data into a conditional distribution of rank p that contains information about the fixed effects and a marginal distribution of rank $(n-p)$ that contains information only about the dispersion parameters (e.g. Verbyla, 1990). The latter distribution is the basis of the restricted likelihood. The partial Laplace method (Taylor, 2005 and Taylor & Verbyla, 2006) produces an approximation to the restricted likelihood. Both an approximation to the unrestricted and restricted likelihoods are derived in this section.

In the derivation for both likelihoods the random normal effects are integrated out prior to the LASSO effects.

$$\begin{aligned}\int_{\mathbb{R}^r} f(\boldsymbol{y},\boldsymbol{u},\boldsymbol{\beta})\partial\boldsymbol{u} &= \int_{\mathbb{R}^r} f(\boldsymbol{y}|\boldsymbol{u},\boldsymbol{\beta})f(\boldsymbol{u})\partial\boldsymbol{u}f(\boldsymbol{\beta})\\ &= f(\boldsymbol{y}|\boldsymbol{\beta})f(\boldsymbol{\beta})\\ &= \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}(2\phi)^q}|\boldsymbol{H}|^{-\frac{1}{2}}\times\\ &\qquad \exp\left(-\frac{1}{2\sigma^2}(\boldsymbol{y}-\boldsymbol{X\tau}-\boldsymbol{L\beta})^T\boldsymbol{H}^{-1}(\boldsymbol{y}-\boldsymbol{X\tau}-\boldsymbol{L\beta})-\frac{1}{\phi}\|\boldsymbol{\beta}\|_1\right)\end{aligned}$$

where $\boldsymbol{H} = \boldsymbol{ZGZ}^T + \boldsymbol{R}$. That is, the distribution of the outcomes marginal to $\boldsymbol{u}$ but conditional on $\boldsymbol{\beta}$ is normal.

8.3.1 *Laplace Approximation*

The Laplace approximation (Chapter 3) provides a convenient method for approximating integrals and it is used here to approximate the unrestricted likelihood.

Theorem 8.1. *The approximate marginal likelihood obtained using the Laplace approximation for the LLMM is*

$$\begin{aligned}\ell(\boldsymbol{\tau},\sigma^2,\phi,\boldsymbol{\gamma},\boldsymbol{\theta}) = -\frac{(n-\kappa)}{2}\log\sigma^2 - q\text{log}\,(2\phi) - \frac{1}{2}\log|\boldsymbol{H}| + \frac{1}{2}\text{log}\,(k_1\ldots k_\kappa)\\ -\frac{1}{2\sigma^2}\left((\boldsymbol{y}-\boldsymbol{X\tau})^T\boldsymbol{H}^{-1}(\boldsymbol{y}-\boldsymbol{X\tau})-\check{\boldsymbol{\beta}}^T\boldsymbol{LH}^{-1}\boldsymbol{L}\check{\boldsymbol{\beta}}\right)\end{aligned}$$

where $\kappa = \text{rank}(\boldsymbol{L}^T\boldsymbol{H}^{-1}\boldsymbol{L})^-$, k_i *is the* i^{th} *non-zero eigenvalue of* $(\boldsymbol{L}^T\boldsymbol{H}^{-1}\boldsymbol{L})^-$ *and* $\check{\boldsymbol{\beta}}$ *is the solution to*

$$\boldsymbol{L}^T\boldsymbol{H}^{-1}\boldsymbol{L}\check{\boldsymbol{\beta}} = \boldsymbol{L}^T\boldsymbol{H}^{-1}(\boldsymbol{y}-\boldsymbol{X\tau}) - \frac{\sigma^2}{\phi}\check{\boldsymbol{v}}.$$

Proof. Denote the exponent of the joint distribution of the outcomes and the LASSO effects, marginal to the random normal effects, by $-h(\boldsymbol{y},\boldsymbol{\beta})$. The sub-differential of $h(\boldsymbol{y},\boldsymbol{\beta})$ is

$$\partial_{\boldsymbol{\beta}}h(\boldsymbol{y},\boldsymbol{\beta}) = -\frac{1}{\sigma^2}\boldsymbol{L}^T\boldsymbol{H}^{-1}(\boldsymbol{y}-\boldsymbol{X\tau}-\boldsymbol{L\beta}) + \frac{1}{\phi}\boldsymbol{v}$$

and the sub-differential's derivative is

$$\frac{\partial(\partial_{\boldsymbol{\beta}}h(\boldsymbol{y},\boldsymbol{\beta}))}{\partial\boldsymbol{\beta}} = \frac{1}{\sigma^2}\boldsymbol{L}^T\boldsymbol{H}^{-1}\boldsymbol{L}.$$

The minimiser of $h(\boldsymbol{y},\boldsymbol{\beta})$ is $\check{\boldsymbol{\beta}}$ that solves $\partial_{\boldsymbol{\beta}}h(\boldsymbol{y},\boldsymbol{\beta}) = 0$; that is, the solution to

$$\boldsymbol{L}^T\boldsymbol{H}^{-1}\boldsymbol{L}\check{\boldsymbol{\beta}} = \boldsymbol{L}^T\boldsymbol{H}^{-1}(\boldsymbol{y}-\boldsymbol{X\tau}) - \frac{\sigma^2}{\phi}\check{\boldsymbol{v}}.$$

Mirroring the proof of Theorem 7.1 a series expansion based on a Taylor series is employed, with the first and second derivative replaced with the sub-differential and its derivative. The expansion is taken around the $\check{\boldsymbol{\beta}}$, the maximiser of $h(\boldsymbol{y},\boldsymbol{\beta})$. This gives

$$\begin{aligned}
h(\boldsymbol{y},\boldsymbol{\beta}) &\approx \frac{1}{2\sigma^2}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\tau}-\boldsymbol{L}\check{\boldsymbol{\beta}})^T\boldsymbol{H}^{-1}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\tau}-\boldsymbol{L}\check{\boldsymbol{\beta}})+\frac{1}{\phi}\check{\boldsymbol{\beta}}^T\check{\boldsymbol{v}}\\
&\qquad+\frac{1}{2\sigma^2}(\boldsymbol{\beta}-\check{\boldsymbol{\beta}})^T\boldsymbol{L}^T\boldsymbol{H}^{-1}\boldsymbol{L}(\boldsymbol{\beta}-\check{\boldsymbol{\beta}})\\
&=\frac{1}{2\sigma^2}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\tau})^T\boldsymbol{H}^{-1}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\tau})+\frac{1}{2\sigma^2}\check{\boldsymbol{\beta}}^T\boldsymbol{L}\boldsymbol{H}^{-1}\boldsymbol{L}\check{\boldsymbol{\beta}}\\
&\qquad-\frac{1}{\sigma^2}\check{\boldsymbol{\beta}}^T\left(\boldsymbol{L}^T\boldsymbol{H}^{-1}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\tau})+\frac{\sigma^2}{\phi}\check{\boldsymbol{v}}\right)+\frac{1}{2\sigma^2}(\boldsymbol{\beta}-\check{\boldsymbol{\beta}})^T\boldsymbol{L}^T\boldsymbol{H}^{-1}\boldsymbol{L}(\boldsymbol{\beta}-\check{\boldsymbol{\beta}})\\
&=\frac{1}{2\sigma^2}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\tau})^T\boldsymbol{H}^{-1}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\tau})-\frac{1}{2\sigma^2}\check{\boldsymbol{\beta}}^T\boldsymbol{L}\boldsymbol{H}^{-1}\boldsymbol{L}\check{\boldsymbol{\beta}}\\
&\qquad+\frac{1}{2\sigma^2}(\boldsymbol{\beta}-\check{\boldsymbol{\beta}})^T\boldsymbol{L}^T\boldsymbol{H}^{-1}\boldsymbol{L}(\boldsymbol{\beta}-\check{\boldsymbol{\beta}}).
\end{aligned}$$

The marginal distribution is

$$\begin{aligned}
f(\boldsymbol{y}) &\approx \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}(2\phi)^q}|\boldsymbol{H}|^{-\frac{1}{2}}(2\pi\sigma^2)^{\frac{\kappa}{2}}(k_1\ldots k_\kappa)^{\frac{1}{2}}\times\\
&\qquad\exp\left(-\frac{1}{2\sigma^2}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\tau})^T\boldsymbol{H}^{-1}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\tau})+\frac{1}{2\sigma^2}\check{\boldsymbol{\beta}}^T\boldsymbol{L}^T\boldsymbol{H}^{-1}\boldsymbol{L}\check{\boldsymbol{\beta}}\right)\times\\
&\qquad\int_{\mathbb{R}^q}\frac{(k_1\ldots k_\kappa)^{-\frac{1}{2}}}{(2\pi\sigma^2)^{\frac{1}{2}}}\exp\left(-\frac{1}{2\sigma^2}(\boldsymbol{\beta}-\check{\boldsymbol{\beta}})^T\boldsymbol{L}^T\boldsymbol{H}^{-1}\boldsymbol{L}(\boldsymbol{\beta}-\check{\boldsymbol{\beta}})\right)\partial\boldsymbol{\beta}\\
&=\frac{1}{(2\pi\sigma^2)^{\frac{n-\kappa}{2}}(2\phi)^q}|\boldsymbol{H}|^{-\frac{1}{2}}(k_1\ldots k_\kappa)^{\frac{1}{2}}\times\\
&\qquad\exp\left(-\frac{1}{2\sigma^2}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\tau})^T\boldsymbol{H}^{-1}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\tau})+\frac{1}{2\sigma^2}\check{\boldsymbol{\beta}}^T\boldsymbol{L}^T\boldsymbol{H}^{-1}\boldsymbol{L}\check{\boldsymbol{\beta}}\right)
\end{aligned}$$

where the second equality holds due to the definition of a singular multivariate normal distribution (see Definition A.1). The theorem is completed by taking the logarithm and disregarding constants. □

8.3.2 *Partial Laplace Approximation*

In Chapter 3 the partial Laplace approximation was reviewed. It was seen that the part of the likelihood free of fixed effects (the restricted likelihood) can be obtained by taking a linear transformation of the outcomes, ignoring the part of the data dependent on fixed effects and then integrating out the random effects. For the LASSO likelihood this will now be employed for an approximate restricted likelihood.

Theorem 8.2. *The approximate restricted marginal likelihood obtained from the partial*

Laplace approximation for the LLMM is

$$\ell_r(\sigma^2, \phi, \boldsymbol{\gamma}, \boldsymbol{\theta}) = -\frac{n-p-\kappa}{2}\log\sigma^2 - q\log\phi - \frac{1}{2}\log|\boldsymbol{H}| - \frac{1}{2}\log|\boldsymbol{X}^T\boldsymbol{H}^{-1}\boldsymbol{X}| + \frac{1}{2}\log(k_1 \dots k_\kappa) - \frac{1}{2\sigma^2}\left(\boldsymbol{y}^T\boldsymbol{P}\boldsymbol{y} - \bar{\boldsymbol{\beta}}^T\boldsymbol{L}^T\boldsymbol{P}\boldsymbol{L}\bar{\boldsymbol{\beta}}\right)$$

where $\kappa = \operatorname{rank}\left((\boldsymbol{L}^T\boldsymbol{P}\boldsymbol{L})^-\right)$, k_i *is the* i^{th} *non-zero eigenvalue of* $(\boldsymbol{L}^T\boldsymbol{P}\boldsymbol{L})^-$ *and* $\bar{\boldsymbol{\beta}}$ *is defined in* (8.2.3)*.*

Proof. Following Verbyla (1990) define $\boldsymbol{W} = [\boldsymbol{W}_1, \boldsymbol{W}_2]$ such that $\boldsymbol{W}_1^T\boldsymbol{X} = \boldsymbol{I}_p$, $\boldsymbol{W}_2^T\boldsymbol{X} = \boldsymbol{0}$ and let

$$\boldsymbol{W}^T\boldsymbol{y} = \begin{pmatrix}\boldsymbol{W}_1^T\boldsymbol{y}\\ \boldsymbol{W}_2^T\boldsymbol{y}\end{pmatrix} = \begin{pmatrix}\boldsymbol{y}_1\\ \boldsymbol{y}_2\end{pmatrix}.$$

Now, $\boldsymbol{y}_1$ depends on fixed effects and $\boldsymbol{y}_2$ does not. The latter is used for estimation of dispersion parameters.

$$\begin{pmatrix}\boldsymbol{y}_1\\ \boldsymbol{y}_2\end{pmatrix}\Bigg|\boldsymbol{\beta} \sim \mathrm{N}\left(\begin{pmatrix}\boldsymbol{\tau} + \boldsymbol{W}_1^T\boldsymbol{L}\boldsymbol{\beta}\\ \boldsymbol{W}_2^T\boldsymbol{L}\boldsymbol{\beta}\end{pmatrix}, \sigma^2\begin{pmatrix}\boldsymbol{W}_1^T\boldsymbol{H}\boldsymbol{W}_1 & \boldsymbol{W}_1^T\boldsymbol{H}\boldsymbol{W}_2\\ \boldsymbol{W}_2^T\boldsymbol{H}\boldsymbol{W}_1 & \boldsymbol{W}_2^T\boldsymbol{H}\boldsymbol{W}_2\end{pmatrix}\right)$$

For the partial Laplace approximation only, consider

$$f(\boldsymbol{y}_2|\boldsymbol{\beta})f(\boldsymbol{\beta}) = \frac{1}{(2\pi\sigma^2)^{\frac{(n-p)}{2}}}\frac{1}{(2\phi)^q}|\boldsymbol{W}_2^T\boldsymbol{H}\boldsymbol{W}_2|^{-\frac{1}{2}}\times \exp\left(-\frac{1}{2\sigma^2}(\boldsymbol{y}_2 - \boldsymbol{W}_2^T\boldsymbol{L}\boldsymbol{\beta})^T(\boldsymbol{W}_2^T\boldsymbol{H}\boldsymbol{W}_2)^{-1}(\boldsymbol{y}_2 - \boldsymbol{W}_2^T\boldsymbol{L}\boldsymbol{\beta}) - \frac{1}{\phi}\|\boldsymbol{\beta}\|_1\right).$$

This can be re-expressed in a form where the transformation does not involve the parameters. Using Theorem A.2 the joint distribution is

$$f(\boldsymbol{y}_2|\boldsymbol{\beta})f(\boldsymbol{\beta}) = \frac{1}{(2\pi\sigma^2)^{\frac{(n-p)}{2}}}\frac{1}{(2\phi)^q}|\boldsymbol{W}^T\boldsymbol{W}|^{-\frac{1}{2}}|\boldsymbol{H}|^{-\frac{1}{2}}|\boldsymbol{X}^T\boldsymbol{H}^{-1}\boldsymbol{X}|^{-\frac{1}{2}}\times \exp\left(-\frac{1}{2\sigma^2}(\boldsymbol{y} - \boldsymbol{L}\boldsymbol{\beta})^T\boldsymbol{P}(\boldsymbol{y} - \boldsymbol{L}\boldsymbol{\beta}) - \frac{1}{\phi}\|\boldsymbol{\beta}\|_1\right).$$

A series expansion, based on a Taylor series, is now used. It is similar to the one in the proof of Theorem 7.1 where the first and second derivatives were replaced by the sub-differential and its derivative. The series expansion is taken around its minimum $\bar{\boldsymbol{\beta}}$, which is the same quantity as the LASSO prediction from the LMME. With the exponent written as $-m_2(\boldsymbol{\beta})$ the sub-differential and its derivative are

$$\partial_{\boldsymbol{\beta}} m_2(\boldsymbol{\beta}) = -\frac{1}{\sigma^2}\boldsymbol{L}^T\boldsymbol{P}(\boldsymbol{y} - \boldsymbol{L}\boldsymbol{\beta}) + \frac{1}{\phi}\boldsymbol{v} \qquad \text{and}$$

$$\frac{\partial(\partial_{\boldsymbol{\beta}} m_2(\boldsymbol{\beta}))}{\partial\boldsymbol{\beta}} = \frac{1}{\sigma^2}\boldsymbol{L}^T\boldsymbol{P}\boldsymbol{L}$$

respectively. The approximation to $m_2(\boldsymbol{\beta})$ is

$$\begin{aligned}
m_2(\boldsymbol{\beta}) &\approx \frac{1}{2\sigma^2}(\boldsymbol{y} - \boldsymbol{L}\bar{\boldsymbol{\beta}})^T\boldsymbol{P}(\boldsymbol{y} - \boldsymbol{L}\bar{\boldsymbol{\beta}}) + \frac{1}{\phi}\|\bar{\boldsymbol{\beta}}\|_1 + \frac{1}{2\sigma^2}(\boldsymbol{\beta} - \bar{\boldsymbol{\beta}})^T\boldsymbol{L}^T\boldsymbol{P}\boldsymbol{L}(\boldsymbol{\beta} - \bar{\boldsymbol{\beta}}) \\
&= \frac{1}{2\sigma^2}\boldsymbol{y}^T\boldsymbol{P}\boldsymbol{y} - \frac{1}{\sigma^2}\bar{\boldsymbol{\beta}}^T\boldsymbol{L}^T\boldsymbol{P}\boldsymbol{y} + \frac{1}{2\sigma^2}\bar{\boldsymbol{\beta}}^T\boldsymbol{L}^T\boldsymbol{P}\boldsymbol{L}\bar{\boldsymbol{\beta}} + \frac{1}{\phi} + \frac{1}{2\sigma^2}(\boldsymbol{\beta} - \bar{\boldsymbol{\beta}})^T\boldsymbol{L}^T\boldsymbol{P}\boldsymbol{L}(\boldsymbol{\beta} - \bar{\boldsymbol{\beta}}) \\
&= \frac{1}{2\sigma^2}\boldsymbol{y}^T\boldsymbol{P}\boldsymbol{y} + \frac{1}{2\sigma^2}\bar{\boldsymbol{\beta}}^T\boldsymbol{L}^T\boldsymbol{P}\boldsymbol{L}\bar{\boldsymbol{\beta}} - \frac{1}{\sigma^2}\bar{\boldsymbol{\beta}}^T\left(\boldsymbol{L}^T\boldsymbol{P}\boldsymbol{y} - \frac{\sigma^2}{\phi}\bar{\boldsymbol{v}}\right) \\
&\qquad\qquad + \frac{1}{2\sigma^2}(\boldsymbol{\beta} - \bar{\boldsymbol{\beta}})^T\boldsymbol{L}^T\boldsymbol{P}\boldsymbol{L}(\boldsymbol{\beta} - \bar{\boldsymbol{\beta}}) \\
&= \frac{1}{2\sigma^2}\boldsymbol{y}^T\boldsymbol{P}\boldsymbol{y} - \frac{1}{2\sigma^2}\bar{\boldsymbol{\beta}}^T\boldsymbol{L}^T\boldsymbol{P}\boldsymbol{L}\bar{\boldsymbol{\beta}} + \frac{1}{2\sigma^2}(\boldsymbol{\beta} - \bar{\boldsymbol{\beta}})^T\boldsymbol{L}^T\boldsymbol{P}\boldsymbol{L}(\boldsymbol{\beta} - \bar{\boldsymbol{\beta}}).
\end{aligned}$$

Using this approximation to the exponent gives the partial Laplace method

$$\begin{aligned}
f(\boldsymbol{y}_2) &= \int_{\mathbb{R}^q} f(\boldsymbol{y}_2|\boldsymbol{\beta})f(\boldsymbol{\beta})\partial\boldsymbol{\beta} \\
&= \frac{1}{(2\pi\sigma^2)^{\frac{(n-p)}{2}}}\frac{1}{(2\phi)^q}|\boldsymbol{W}^T\boldsymbol{W}|^{-\frac{1}{2}}|\boldsymbol{H}|^{-\frac{1}{2}}|\boldsymbol{X}^T\boldsymbol{H}^{-1}\boldsymbol{X}|^{-\frac{1}{2}}\times \\
&\qquad \exp\left(-\frac{1}{2\sigma^2}(\boldsymbol{y}^T\boldsymbol{P}\boldsymbol{y} - \bar{\boldsymbol{\beta}}^T\boldsymbol{L}^T\boldsymbol{P}\boldsymbol{L}\bar{\boldsymbol{\beta}})\right)(k_1\ldots k_\kappa)^{\frac{1}{2}}(2\pi\sigma^2)^{\frac{\kappa}{2}}\times \\
&\qquad \int_{\mathbb{R}^q}\frac{1}{(2\phi\sigma^2)^{\frac{\kappa}{2}}}(k_1\ldots k_\kappa)^{-\frac{1}{2}}\exp\left(-\frac{1}{2\sigma^2}(\boldsymbol{\beta} - \bar{\boldsymbol{\beta}})^T\boldsymbol{L}^T\boldsymbol{P}\boldsymbol{L}(\boldsymbol{\beta} - \bar{\boldsymbol{\beta}})\right)\partial\boldsymbol{\beta} \\
&= \frac{1}{(2\pi\sigma^2)^{\frac{(n-p-\kappa)}{2}}}\frac{1}{(2\phi)^q}|\boldsymbol{W}^T\boldsymbol{W}|^{-\frac{1}{2}}|\boldsymbol{H}|^{-\frac{1}{2}}|\boldsymbol{X}^T\boldsymbol{H}^{-1}\boldsymbol{X}|^{-\frac{1}{2}}(k_1\ldots k_\kappa)^{\frac{1}{2}}\times \\
&\qquad \exp\left(-\frac{1}{2\sigma^2}(\boldsymbol{y}^T\boldsymbol{P}\boldsymbol{y} - \bar{\boldsymbol{\beta}}^T\boldsymbol{L}^T\boldsymbol{P}\boldsymbol{L}\bar{\boldsymbol{\beta}})\right)
\end{aligned}$$

where the third equality holds because the integrand is the functional form for a singular multivariate normal distribution (see Definition A.1). Taking the logarithm and disregarding constants gives the approximate restricted likelihood. □

The differences between the approximate marginal distribution and the approximate restricted marginal distribution are similar to the differences observed for the standard normal mixed model. That is the matrix $\boldsymbol{H}^{-1}$ in the unrestricted likelihood is replaced by $\boldsymbol{P}$, the term $-\frac{1}{2}\log|\boldsymbol{X}^T\boldsymbol{H}^{-1}\boldsymbol{X}|$ is included and the coefficient of $\log\sigma^2$ is $(n-p-\kappa)/2$ rather than $(n-\kappa)/2$. The approximate unrestricted likelihood requires knowledge of the unknown parameter $\boldsymbol{\tau}$. Also, the estimate $\check{\boldsymbol{\beta}}$ is not that obtained from the LMME (8.2.1) and it depends on the unknown $\boldsymbol{\tau}$. Usually, a plug-in estimator of $\boldsymbol{\tau}$ would be used when calculating the likelihood. However, this is circumvented in the approximate restricted likelihood and the estimates $\bar{\boldsymbol{\beta}}$ are exactly those that solve the LMME. Given that these differences occur, it seems that the approximate restricted likelihood is more appropriate. The remainder of this chapter will use only this likelihood.

Under the condition that $\kappa = q < n - p$, the approximate restricted likelihood has a particular form. This form is similar to the exact restricted likelihood for the standard mixed models.

Corollary 8.1. *If $\boldsymbol{L}^T\boldsymbol{P}\boldsymbol{L}$ is non-singular then the approximate restricted likelihood is*

$$\ell_r(\sigma^2,\phi,\boldsymbol{\gamma},\boldsymbol{\theta}) = -\frac{n-p-q}{2}\log\sigma^2 - q\log\phi - \frac{1}{2}\log|\boldsymbol{H}| - \frac{1}{2}\log|\boldsymbol{X}^T\boldsymbol{H}^{-1}\boldsymbol{X}| + \frac{1}{2}\log|\boldsymbol{L}^T\boldsymbol{P}\boldsymbol{L}| - \frac{1}{2\sigma^2}\left(\boldsymbol{y}^T\boldsymbol{P}\boldsymbol{y} - \bar{\boldsymbol{\beta}}^T\boldsymbol{L}^T\boldsymbol{P}\boldsymbol{L}\bar{\boldsymbol{\beta}}\right).$$

Proof. Use Theorem 8.2 and note that $(k_1\dots k_\kappa) = |\boldsymbol{L}^T\boldsymbol{P}\boldsymbol{L}|$. □

8.4 Derivatives of LASSO Estimates

Obtaining the score equations and the information for the approximate restricted likelihood requires the derivatives of the LASSO estimates. The derivatives w.r.t. σ^2 and ϕ are generalisations of those for the LASSO random effects model in Theorem 7.2. The derivation of those derivatives (see Chapter 7) was enabled by noticing that the derivatives of sub-vectors of $\bar{\boldsymbol{\beta}}$ and $\bar{\boldsymbol{v}}$ are zero. This argument is used again. Also, Lemma 7.1 is needed in slight generalisation. This is presented first.

Lemma 8.1. *Define the $q\times q$ matrix $\boldsymbol{D}$ as*

$$\boldsymbol{L}^T\boldsymbol{P}\boldsymbol{L} = \boldsymbol{D} = \begin{pmatrix}\boldsymbol{D}_{\mathcal{S}\mathcal{S}} & \boldsymbol{D}_{\mathcal{S}\mathcal{Z}}\\ \boldsymbol{D}_{\mathcal{Z}\mathcal{S}} & \boldsymbol{D}_{\mathcal{Z}\mathcal{Z}}\end{pmatrix} \quad \textit{and} \quad (\boldsymbol{L}^T\boldsymbol{P}\boldsymbol{L})^- = \boldsymbol{D}^- = \begin{pmatrix}\boldsymbol{D}^{\mathcal{S}\mathcal{S}} & \boldsymbol{D}^{\mathcal{S}\mathcal{Z}}\\ \boldsymbol{D}^{\mathcal{Z}\mathcal{S}} & \boldsymbol{D}^{\mathcal{Z}\mathcal{Z}}\end{pmatrix}$$

where $\boldsymbol{L}$ is a $n\times q$ matrix and q may be larger than n and $\boldsymbol{P}$ is a $n\times n$ symmetric matrix, then

$$\left(\boldsymbol{I}_{\mathcal{S}}, -\boldsymbol{D}^{\mathcal{S}\mathcal{Z}}\left(\boldsymbol{D}^{\mathcal{Z}\mathcal{Z}}\right)^-\right)\boldsymbol{D}^- = \begin{pmatrix}\boldsymbol{D}^-_{\mathcal{S}\mathcal{S}} & \boldsymbol{0}\end{pmatrix}.$$

Proof. The proof follows that of Lemma 7.1 with $\boldsymbol{D}$ redefined. □

Theorem 8.3. *Let $\boldsymbol{\eta}^T = (\boldsymbol{\gamma}^T, \boldsymbol{\theta}^T)$ whose i^{th} element is η_i. The derivatives of the LASSO estimates with respect to σ^2, ϕ and η_i are*

$$\frac{\partial\bar{\boldsymbol{\beta}}}{\partial\sigma^2} = -\frac{1}{\phi}\begin{pmatrix}(\boldsymbol{L}_{\mathcal{S}}^T\boldsymbol{P}\boldsymbol{L}_{\mathcal{S}})^{-1} & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{0}\end{pmatrix}\bar{\boldsymbol{v}},$$

$$\frac{\partial\bar{\boldsymbol{\beta}}}{\partial\phi} = -\frac{\sigma^2}{\phi^2}\begin{pmatrix}(\boldsymbol{L}_{\mathcal{S}}^T\boldsymbol{P}\boldsymbol{L}_{\mathcal{S}})^{-1} & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{0}\end{pmatrix}\bar{\boldsymbol{v}} \qquad \textit{and}$$

$$\frac{\partial\bar{\boldsymbol{\beta}}}{\partial\eta_i} = -\begin{pmatrix}(\boldsymbol{L}_{\mathcal{S}}^T\boldsymbol{P}\boldsymbol{L}_{\mathcal{S}})^{-1} & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{0}\end{pmatrix}\boldsymbol{L}^T\boldsymbol{P}\dot{\boldsymbol{H}}_i\boldsymbol{P}(\boldsymbol{y}-\boldsymbol{L}\bar{\boldsymbol{\beta}})$$

respectively where $\dot{\boldsymbol{H}}_i$ is the derivative of $\boldsymbol{H}$ w.r.t η_i.

Proof. The derivatives w.r.t. σ^2 and ϕ are trivial extensions of Theorem 7.2. The derivative w.r.t. η_i is now presented. Denote the least squares estimate of the LASSO effects as $\hat{\boldsymbol{\beta}}$.

Then

$$\begin{aligned} \frac{\partial \bar{\boldsymbol{\beta}}_{\mathcal{Z}}}{\partial \eta_i} &= \mathbf{0} = \frac{\partial \hat{\boldsymbol{\beta}}_{\mathcal{Z}}}{\partial \eta_i} - \frac{\sigma^2}{\phi}\frac{\partial \boldsymbol{D}^{\mathcal{Z}\cdot}}{\partial \eta_i}\bar{\boldsymbol{v}} - \frac{\sigma^2}{\phi}\boldsymbol{D}^{\mathcal{Z}\mathcal{S}}\frac{\partial \bar{\boldsymbol{v}}_{\mathcal{S}}}{\partial \eta_i} - \frac{\sigma^2}{\phi}\boldsymbol{D}^{\mathcal{Z}\mathcal{Z}}\frac{\partial \bar{\boldsymbol{v}}_{\mathcal{Z}}}{\partial \eta_i} \\ \therefore \qquad \frac{\partial \bar{\boldsymbol{v}}_{\mathcal{Z}}}{\partial \eta_i} &= \frac{\phi}{\sigma^2}\left(\boldsymbol{D}^{\mathcal{Z}\mathcal{Z}}\right)^{-}\frac{\partial \hat{\boldsymbol{\beta}}_{\mathcal{Z}}}{\partial \eta_i} - \left(\boldsymbol{D}^{\mathcal{Z}\mathcal{Z}}\right)^{-}\frac{\partial \boldsymbol{D}^{\mathcal{Z}\cdot}}{\partial \eta_i}\bar{\boldsymbol{v}} \end{aligned}$$

as $\partial \bar{\boldsymbol{v}}_{\mathcal{S}}/\partial \eta_i = \mathbf{0}$. Now

$$\begin{aligned} \frac{\partial \bar{\boldsymbol{\beta}}_{\mathcal{S}}}{\partial \eta_i} &= \frac{\partial \hat{\boldsymbol{\beta}}_{\mathcal{S}}}{\partial \eta_i} - \frac{\sigma^2}{\phi}\frac{\partial \boldsymbol{D}^{\mathcal{S}\cdot}}{\partial \eta_i}\bar{\boldsymbol{v}} - \frac{\sigma^2}{\phi}\boldsymbol{D}^{\mathcal{S}\mathcal{S}}\frac{\partial \bar{\boldsymbol{v}}_{\mathcal{S}}}{\partial \eta_i} - \frac{\sigma^2}{\phi}\boldsymbol{D}^{\mathcal{S}\mathcal{Z}}\frac{\partial \bar{\boldsymbol{v}}_{\mathcal{Z}}}{\partial \eta_i} \\ &= \frac{\partial \hat{\boldsymbol{\beta}}_{\mathcal{S}}}{\partial \eta_i} - \frac{\sigma^2}{\phi}\frac{\partial \boldsymbol{D}^{\mathcal{S}\cdot}}{\partial \eta_i}\bar{\boldsymbol{v}} - \boldsymbol{D}^{\mathcal{S}\mathcal{Z}}\left(\boldsymbol{D}^{\mathcal{Z}\mathcal{Z}}\right)^{-}\frac{\partial \hat{\boldsymbol{\beta}}_{\mathcal{Z}}}{\partial \eta_i} + \frac{\sigma^2}{\phi}\boldsymbol{D}^{\mathcal{S}\mathcal{Z}}\left(\boldsymbol{D}^{\mathcal{Z}\mathcal{Z}}\right)^{-}\frac{\partial \boldsymbol{D}^{\mathcal{Z}\cdot}}{\partial \eta_i}\bar{\boldsymbol{v}} \\ &= \left(\boldsymbol{I}_{\mathcal{S}}, -\boldsymbol{D}^{\mathcal{S}\mathcal{Z}}\left(\boldsymbol{D}^{\mathcal{Z}\mathcal{Z}}\right)^{-}\right)\frac{\partial \hat{\boldsymbol{\beta}}}{\partial \eta_i} - \frac{\sigma^2}{\phi}\left(\boldsymbol{I}_{\mathcal{S}}, -\boldsymbol{D}^{\mathcal{S}\mathcal{Z}}\left(\boldsymbol{D}^{\mathcal{Z}\mathcal{Z}}\right)^{-}\right)\frac{\partial \boldsymbol{D}^{-}}{\partial \eta_i}\bar{\boldsymbol{v}} \\ &= \left(\boldsymbol{I}_{\mathcal{S}}, -\boldsymbol{D}^{\mathcal{S}\mathcal{Z}}\left(\boldsymbol{D}^{\mathcal{Z}\mathcal{Z}}\right)^{-}\right)\left(-\boldsymbol{D}^{-}\frac{\partial \boldsymbol{D}}{\partial \eta_i}\boldsymbol{D}^{-}\boldsymbol{L}^T\boldsymbol{P}\boldsymbol{y} + \boldsymbol{D}^{-}\boldsymbol{L}^T\frac{\partial \boldsymbol{P}}{\partial \eta_i}\boldsymbol{y} + \frac{\sigma^2}{\phi}\boldsymbol{D}^{-}\frac{\partial \boldsymbol{D}}{\partial \eta_i}\boldsymbol{D}^{-}\bar{\boldsymbol{v}}\right) \\ &= \left(\boldsymbol{I}_{\mathcal{S}}, -\boldsymbol{D}^{\mathcal{S}\mathcal{Z}}\left(\boldsymbol{D}^{\mathcal{Z}\mathcal{Z}}\right)^{-}\right)\Big(\boldsymbol{D}^{-}\boldsymbol{L}^T\boldsymbol{P}\dot{\boldsymbol{H}}_i\boldsymbol{P}\boldsymbol{L}\boldsymbol{D}^{-}\boldsymbol{L}^T\boldsymbol{P}\boldsymbol{y} - \boldsymbol{D}^{-}\boldsymbol{L}^T\boldsymbol{P}\dot{\boldsymbol{H}}_i\boldsymbol{P}\boldsymbol{y} \\ &\qquad\qquad\qquad\qquad - \frac{\sigma^2}{\phi}\boldsymbol{D}^{-}\boldsymbol{L}^T\boldsymbol{P}\dot{\boldsymbol{H}}_i\boldsymbol{P}\boldsymbol{L}\boldsymbol{D}^{-}\bar{\boldsymbol{v}}\Big) \\ &= -\left(\boldsymbol{I}_{\mathcal{S}}, -\boldsymbol{D}^{\mathcal{S}\mathcal{Z}}\left(\boldsymbol{D}^{\mathcal{Z}\mathcal{Z}}\right)^{-}\right)\boldsymbol{D}^{-}\boldsymbol{L}^T\boldsymbol{P}\dot{\boldsymbol{H}}_i\boldsymbol{P}\left(\boldsymbol{y} - \boldsymbol{L}\left(\boldsymbol{D}^{-}\boldsymbol{L}^T\boldsymbol{P}\boldsymbol{y} - \frac{\sigma^2}{\phi}\boldsymbol{D}^{-}\bar{\boldsymbol{v}}\right)\right). \end{aligned}$$

Using Theorem 8.1 and the form of the LASSO estimates in (8.2.3) this derivative is

$$\frac{\partial \bar{\boldsymbol{\beta}}_{\mathcal{S}}}{\partial \eta_i} = -\left(\boldsymbol{D}_{\mathcal{S}\mathcal{S}}^{-}, \mathbf{0}\right)\boldsymbol{L}^T\boldsymbol{P}\dot{\boldsymbol{H}}_i\boldsymbol{P}\left(\boldsymbol{y} - \boldsymbol{L}\bar{\boldsymbol{\beta}}\right).$$

Hence

$$\frac{\partial \bar{\boldsymbol{\beta}}}{\partial \gamma_i} = -\begin{pmatrix} (\boldsymbol{L}_{\mathcal{S}}^T\boldsymbol{P}\boldsymbol{L}_{\mathcal{S}})^{-} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}\boldsymbol{L}^T\boldsymbol{P}\dot{\boldsymbol{H}}_i\boldsymbol{P}(\boldsymbol{y} - \boldsymbol{L}\bar{\boldsymbol{\beta}}).$$

In all cases where $q_{\mathcal{S}} \leq n$ the matrix $\boldsymbol{L}_{\mathcal{S}}^T\boldsymbol{P}\boldsymbol{L}_{\mathcal{S}}$ will be of full rank. Theorem 6 of Osborne et al. (2000) shows that there will be at most n non-zero LASSO predictions. □

8.5 Estimation of Dispersion Parameters

8.5.1 *Unadjusted Score Equations*

Theorem 8.4. *The un-adjusted score equations from the approximate restricted likelihood in Theorem 8.2 are*

$$\begin{aligned} \frac{\partial \ell_r}{\partial \sigma^2} &= -\frac{n-p-\kappa}{2\sigma^2} + \frac{1}{2\sigma^4}\left(\boldsymbol{y}^T\boldsymbol{P}\boldsymbol{y} - \bar{\boldsymbol{\beta}}^T\boldsymbol{L}^T\boldsymbol{P}\boldsymbol{L}\bar{\boldsymbol{\beta}}\right) - \frac{1}{\sigma^2\phi}\|\bar{\boldsymbol{\beta}}\|_1 \\ \frac{\partial \ell_r}{\partial \phi} &= -\frac{q}{\phi} + \frac{1}{\phi^2}\|\bar{\boldsymbol{\beta}}\|_1 \\ \frac{\partial \ell_r}{\partial \eta_i} &= -\frac{1}{2}\mathrm{tr}\left(\boldsymbol{P}\dot{\boldsymbol{H}}_i\right) + \frac{1}{2}\frac{\partial \log\left(k_1 \ldots k_\kappa\right)}{\partial \eta_i} + \frac{1}{2\sigma^2}(\boldsymbol{y} - \boldsymbol{L}\bar{\boldsymbol{\beta}})^T\boldsymbol{L}^T\dot{\boldsymbol{H}}_i\boldsymbol{L}(\boldsymbol{y} - \boldsymbol{L}\bar{\boldsymbol{\beta}}) \end{aligned}$$

Proof. The score equations for σ^2 and ϕ are direct extensions to those presented in Theorem 7.3. Only the derivation for the η_i score is presented. The derivation uses matrix derivatives in Result B.1.

$$\begin{aligned}\frac{\partial \ell_r}{\partial \eta_i} = &-\frac{1}{2}\text{tr}\left(\boldsymbol{H}^{-1}\dot{\boldsymbol{H}}\right) + \frac{1}{2}\text{tr}\left((\boldsymbol{X}^T\boldsymbol{H}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{H}^{-1}\dot{\boldsymbol{H}}_i\boldsymbol{H}^{-1}\boldsymbol{X}\right) + \frac{1}{2}\frac{\partial \log(k_1 \ldots k_\kappa)}{\partial \eta_i} \\ &\frac{1}{2\sigma^2}\boldsymbol{y}^T\boldsymbol{P}\dot{\boldsymbol{H}}_i\boldsymbol{P}\boldsymbol{y} - \frac{1}{\sigma^2}(\boldsymbol{y}-\boldsymbol{L}\bar{\boldsymbol{\beta}})^T\boldsymbol{P}\dot{\boldsymbol{H}}\boldsymbol{P}\boldsymbol{L}\begin{pmatrix}(\boldsymbol{L}^T\boldsymbol{P}\boldsymbol{L})^{-1} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0}\end{pmatrix}\boldsymbol{L}^T\boldsymbol{P}\boldsymbol{L}\bar{\boldsymbol{\beta}} \\ &-\frac{1}{2\sigma^2}\bar{\boldsymbol{\beta}}^T\boldsymbol{L}^T\boldsymbol{P}\dot{\boldsymbol{H}}\boldsymbol{P}\boldsymbol{L}\bar{\boldsymbol{\beta}} \\ = &-\frac{1}{2}\text{tr}\left(\boldsymbol{P}\dot{\boldsymbol{H}}_i\right) + \frac{1}{2}\frac{\partial \log(k_1 \ldots k_\kappa)}{\partial \eta_i} + \frac{1}{2\sigma^2}(\boldsymbol{y}-\boldsymbol{L}\bar{\boldsymbol{\beta}})^T\boldsymbol{P}\dot{\boldsymbol{H}}_i\boldsymbol{P}(\boldsymbol{y}-\boldsymbol{L}\bar{\boldsymbol{\beta}}).\end{aligned}$$

□

Corollary 8.2. *If the matrix $\boldsymbol{L}^T\boldsymbol{P}\boldsymbol{L}$ is non-singular then the score equation for η_i is*

$$\frac{\partial \ell_r}{\partial \eta_i} = -\frac{1}{2}\text{tr}\left(\boldsymbol{P}^*\dot{\boldsymbol{H}}_i\right) + \frac{1}{2\sigma^2}(\boldsymbol{y}-\boldsymbol{L}\bar{\boldsymbol{\beta}})^T\boldsymbol{L}^T\dot{\boldsymbol{H}}_i\boldsymbol{L}(\boldsymbol{y}-\boldsymbol{L}\bar{\boldsymbol{\beta}})$$

where $\boldsymbol{P}^* = \boldsymbol{P} - \boldsymbol{P}\boldsymbol{L}^T(\boldsymbol{L}^T\boldsymbol{P}\boldsymbol{L})^{-1}\boldsymbol{L}\boldsymbol{P}$.

Proof. Consider the approximate restricted likelihood in Corollary 8.1. In particular, consider only the determinant terms (other terms were differentiated in Theorem 8.4). The derivative of the determinant terms is

$$\begin{aligned}\frac{\partial}{\partial \eta_i}&\left(-\frac{1}{2}\log|\boldsymbol{H}| - \frac{1}{2}\log|\boldsymbol{X}^T\boldsymbol{H}^{-1}\boldsymbol{X}| + \frac{1}{2}\log|\boldsymbol{L}^T\boldsymbol{P}\boldsymbol{L}|\right) \\ &= -\frac{1}{2}\text{tr}\left(\boldsymbol{P}\dot{\boldsymbol{H}}_i\right) + \frac{1}{2}\text{tr}\left((\boldsymbol{L}^T\boldsymbol{P}\boldsymbol{L})^{-1}\boldsymbol{L}^T\boldsymbol{P}\dot{\boldsymbol{H}}_i\boldsymbol{P}\boldsymbol{L}\right) \\ &= -\frac{1}{2}\text{tr}\left(\boldsymbol{P}^*\dot{\boldsymbol{H}}_i\right).\end{aligned}$$

□

8.5.2 *Observed Information*

The information for the dispersion parameters will provide a method for solving the score equations for η_i.

Theorem 8.5. *The observed information for the unadjusted scores has elements*

$$
\begin{aligned}
-\frac{\partial^2 \ell_r}{\partial\sigma^2\partial\sigma^2} &= \frac{n-p-\kappa+\frac{4}{\phi}\|\bar{\boldsymbol{\beta}}\|_1}{2\sigma^4} - \frac{1}{\sigma^6}(\boldsymbol{y}^T\boldsymbol{P}\boldsymbol{y} - \bar{\boldsymbol{\beta}}^T\boldsymbol{L}^T\boldsymbol{P}\boldsymbol{L}\bar{\boldsymbol{\beta}}) + \frac{1}{\sigma^2\phi^2}\bar{\boldsymbol{v}}_{\mathcal{S}}^T(\boldsymbol{L}_{\mathcal{S}}^T\boldsymbol{P}\boldsymbol{L}_{\mathcal{S}})^{-1}\bar{\boldsymbol{v}}_{\mathcal{S}},\\
-\frac{\partial^2 \ell_r}{\partial\phi\partial\phi} &= \frac{q}{\phi^2} - \frac{2}{\phi^3}\|\bar{\boldsymbol{\beta}}\|_1 + \frac{\sigma^2}{\phi^4}\bar{\boldsymbol{v}}_{\mathcal{S}}^T(\boldsymbol{L}_{\mathcal{S}}^T\boldsymbol{P}\boldsymbol{L}_{\mathcal{S}})^{-1}\bar{\boldsymbol{v}}_{\mathcal{S}},\\
-\frac{\partial^2 \ell_r}{\partial\eta_i\partial\eta_j} &= \frac{1}{2}\mathrm{tr}\left(\boldsymbol{P}\dot{\boldsymbol{H}}_i\boldsymbol{P}\dot{\boldsymbol{H}}_j\right) - \frac{1}{2}\mathrm{tr}\left(\boldsymbol{P}\ddot{\boldsymbol{H}}_{ij}\right) + \frac{1}{2}\frac{\partial^2\log\left(k_1\ldots k_\kappa\right)}{\partial\eta_i\partial\eta_j}\\
&\quad - \frac{1}{\sigma^2}(\boldsymbol{y}-\boldsymbol{L}\bar{\boldsymbol{\beta}})^T\boldsymbol{P}\dot{\boldsymbol{H}}_i\boldsymbol{T}\dot{\boldsymbol{H}}_j\boldsymbol{P}(\boldsymbol{y}-\boldsymbol{L}\bar{\boldsymbol{\beta}}) + \frac{1}{2\sigma^2}(\boldsymbol{y}-\boldsymbol{L}\bar{\boldsymbol{\beta}})^T\boldsymbol{P}\ddot{\boldsymbol{H}}_{ij}\boldsymbol{P}(\boldsymbol{y}-\boldsymbol{L}\bar{\boldsymbol{\beta}}),\\
-\frac{\partial^2 \ell_r}{\partial\phi\partial\sigma^2} &= -\frac{1}{\phi^2}\bar{\boldsymbol{v}}_{\mathcal{S}}^T(\boldsymbol{L}_{\mathcal{S}}^T\boldsymbol{P}\boldsymbol{L}_{\mathcal{S}})^{-1}\bar{\boldsymbol{v}}_{\mathcal{S}},\\
-\frac{\partial^2 \ell_r}{\partial\eta_i\partial\sigma^2} &= -\frac{1}{2\sigma^4}(\boldsymbol{y}-\boldsymbol{L}\bar{\boldsymbol{\beta}})^T\boldsymbol{P}\dot{\boldsymbol{H}}_i\boldsymbol{P}(\boldsymbol{y}-\boldsymbol{L}\bar{\boldsymbol{\beta}})\\
&\quad + \frac{1}{\sigma^2\phi}\bar{\boldsymbol{v}}_{\mathcal{S}}^T(\boldsymbol{L}_{\mathcal{S}}^T\boldsymbol{P}\boldsymbol{L}_{\mathcal{S}})^{-1}\boldsymbol{L}_{\mathcal{S}}\boldsymbol{P}\dot{\boldsymbol{H}}_i\boldsymbol{P}(\boldsymbol{y}-\boldsymbol{L}\bar{\boldsymbol{\beta}}) \qquad \textit{and}\\
-\frac{\partial^2 \ell_r}{\partial\phi\partial\eta_i} &= -\frac{1}{\phi^2}\bar{\boldsymbol{v}}_{\mathcal{S}}^T(\boldsymbol{L}_{\mathcal{S}}^T\boldsymbol{P}\boldsymbol{L}_{\mathcal{S}})^{-1}\boldsymbol{L}_{\mathcal{S}}^T\boldsymbol{P}\dot{\boldsymbol{H}}_i\boldsymbol{P}(\boldsymbol{y}-\boldsymbol{L}\bar{\boldsymbol{\beta}})
\end{aligned}
$$

where $\boldsymbol{T} = \boldsymbol{P} - \boldsymbol{P}\boldsymbol{L}_{\mathcal{S}}(\boldsymbol{L}_{\mathcal{S}}^T\boldsymbol{P}\boldsymbol{L}_{\mathcal{S}})^{-1}\boldsymbol{L}_{\mathcal{S}}\boldsymbol{P}$, *and* $\ddot{\boldsymbol{H}}_{ij}$ *is the second derivative of* $\boldsymbol{H}$ *with respect to* η_i *and* η_j*.*

Proof. Only the result for $(\partial^2\ell_r)/(\partial\eta_i\partial\eta_j)$ is considered as the others are relatively straightforward. First, consider the derivatives of the quadratic form $Q = (\boldsymbol{y}-\boldsymbol{L}\bar{\boldsymbol{\beta}})^T\boldsymbol{P}\dot{\boldsymbol{H}}_i\boldsymbol{P}(\boldsymbol{y}-\boldsymbol{L}\bar{\boldsymbol{\beta}})/(2\sigma^2)$.

$$
\begin{aligned}
\frac{\partial Q}{\partial\eta_j} &= \frac{1}{\sigma^2}(\boldsymbol{y}-\boldsymbol{L}\bar{\boldsymbol{\beta}})^T\boldsymbol{P}\dot{\boldsymbol{H}}_i\boldsymbol{P}\boldsymbol{L}\begin{pmatrix}(\boldsymbol{L}_{\mathcal{S}}^T\boldsymbol{P}\boldsymbol{L}_{\mathcal{S}})^{-1} & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{0}\end{pmatrix}\boldsymbol{L}^T\boldsymbol{P}\dot{\boldsymbol{H}}_j\boldsymbol{P}(\boldsymbol{y}-\boldsymbol{L}\bar{\boldsymbol{\beta}})\\
&\quad - \frac{1}{\sigma^2}(\boldsymbol{y}-\boldsymbol{L}\bar{\boldsymbol{\beta}})^T\boldsymbol{P}\dot{\boldsymbol{H}}_i\boldsymbol{P}\dot{\boldsymbol{H}}_j\boldsymbol{P}(\boldsymbol{y}-\boldsymbol{L}\bar{\boldsymbol{\beta}}) + \frac{1}{2\sigma^2}(\boldsymbol{y}-\boldsymbol{L}\bar{\boldsymbol{\beta}})^T\boldsymbol{P}\ddot{\boldsymbol{H}}_{ij}\boldsymbol{P}(\boldsymbol{y}-\boldsymbol{L}\bar{\boldsymbol{\beta}})\\
&= -\frac{1}{\sigma^2}(\boldsymbol{y}-\boldsymbol{L}\bar{\boldsymbol{\beta}})^T\boldsymbol{P}\dot{\boldsymbol{H}}_i\left(\boldsymbol{P}-\boldsymbol{P}\boldsymbol{L}_{\mathcal{S}}(\boldsymbol{L}_{\mathcal{S}}^T\boldsymbol{P}\boldsymbol{L}_{\mathcal{S}})^{-1}\boldsymbol{L}_{\mathcal{S}}^T\boldsymbol{P}\right)\dot{\boldsymbol{H}}_j\boldsymbol{P}(\boldsymbol{y}-\boldsymbol{L}\bar{\boldsymbol{\beta}})\\
&\quad + \frac{1}{2\sigma^2}(\boldsymbol{y}-\boldsymbol{L}\bar{\boldsymbol{\beta}})^T\boldsymbol{P}\ddot{\boldsymbol{H}}_{ij}\boldsymbol{P}(\boldsymbol{y}-\boldsymbol{L}\bar{\boldsymbol{\beta}}).
\end{aligned}
$$

The derivatives of the log-determinants are given in the proof of Theorem A.5. □

8.5.3 *Adjusted Score Equations and Estimates*

The score equations for σ^2 and ϕ arising from the random model containing only LASSO effects, presented in Chapter 7, were biased. To overcome this the expected score equations were subtracted from the observed scores in the manner proposed by McCullagh & Tibshirani (1990). The expected scores are calculated using a parametric bootstrap routine. The same method is employed for the LLMM. This adjustment also has the fortuitous effect of removing the need to calculate the derivative $\partial \log\left(k_1 \ldots k_\kappa\right)/\partial\eta_i$.

As with the model containing only LASSO effects, there are closed form estimates of the parameters σ^2 and ϕ. However, there is not a closed form for the dispersion parameters $\boldsymbol{\eta}$ relating to the random normal effects and residuals.

Theorem 8.6. *The estimates for σ^2 and ϕ arising from the adjusted score equations are*

$$\begin{aligned}\hat{\sigma}^2 &= \frac{\boldsymbol{y}^T\boldsymbol{P}\boldsymbol{y} - \bar{\boldsymbol{\beta}}^T\boldsymbol{L}^T\boldsymbol{P}\boldsymbol{L}\bar{\boldsymbol{\beta}}}{n - p - \kappa + \frac{2}{\phi}\|\bar{\boldsymbol{\beta}}\|_1 + 2\hat{\sigma}^2 M_{\sigma^2}}\\ \hat{\phi} &= \frac{\|\bar{\boldsymbol{\beta}}\|_1}{q + \hat{\phi}M_\phi}\end{aligned}$$

where M_{σ^2} and M_ϕ are the (bootstrap) estimates of the expected values for the score equations for σ^2 and ϕ respectively.

Proof. Follows directly from the proof of Theorem 7.5. □

Theorem 8.7. *The adjusted score equation for η_i arising from subtracting a bootstrap estimate of the score's expected value is*

$$\begin{aligned}\left(\frac{\partial\ell(\sigma^2,\phi,\boldsymbol{\eta})}{\partial\eta_i}\right)^* = {}& \frac{1}{2\sigma^2}(\boldsymbol{y} - \boldsymbol{L}\bar{\boldsymbol{\beta}})^T\boldsymbol{P}\dot{\boldsymbol{H}}_i\boldsymbol{P}(\boldsymbol{y} - \boldsymbol{L}\bar{\boldsymbol{\beta}})\\ & - \frac{1}{2\sigma^2 B}\sum_{j=1}^{B}(\boldsymbol{y}_j^* - \boldsymbol{L}\bar{\boldsymbol{\beta}}_j^*)^T\boldsymbol{P}\dot{\boldsymbol{H}}_i\boldsymbol{P}(\boldsymbol{y}_j^* - \boldsymbol{L}\bar{\boldsymbol{\beta}}_j^*)\end{aligned}$$

where $\boldsymbol{y}_j^*$ and $\bar{\boldsymbol{\beta}}_j^*$ are the j^{th} bootstrap sample and LASSO estimate respectively and B is the number of bootstrap samples performed.

Proof. Consider the unadjusted score in Theorem 8.4. The bootstrap adjustment term is the average of the scores of the simulated data. Each of these simulated scores will have a component corresponding to $-\mathrm{tr}\left(\boldsymbol{P}\dot{\boldsymbol{H}}_i\right)$ and $\partial\log\left(k_1 \ldots k_\kappa\right)/\partial\eta_i$ which are constant across all B simulations. Upon subtraction these terms cancel leaving only those terms in the Theorem statement. □

A method of scoring is used to solve the adjusted score equations for $\boldsymbol{\eta}$. The information matrix (negative Hessian) used is based on the adjusted observed information. The observed information is not used as it can be computationally expensive to evaluate. The approximate information used closely mirrors that of the information used in the average information algorithm (Chapter 3) and its elements can be efficiently calculated. The information used in the scoring routine has elements

$$\begin{aligned}
\mathcal{I}(\sigma^2, \phi) &= \mathcal{I}(\eta_i, \phi) = 0 \\
\mathcal{I}(\sigma^2, \sigma^2) &= \frac{(\boldsymbol{y} - \boldsymbol{L}\bar{\boldsymbol{\beta}})^T \boldsymbol{P}(\boldsymbol{y} - \boldsymbol{L}\bar{\boldsymbol{\beta}})}{2\sigma^2} \\
\mathcal{I}(\sigma^2, \eta_i) &= \frac{1}{2\sigma^4}(\boldsymbol{y} - \boldsymbol{L}\bar{\boldsymbol{\beta}})^T \boldsymbol{P}\dot{\boldsymbol{H}}_i\boldsymbol{P}(\boldsymbol{y} - \boldsymbol{L}\bar{\boldsymbol{\beta}}) \\
&\qquad - \frac{1}{2\sigma^4 B}\sum_{j=1}^{B}(\boldsymbol{y}^* - \boldsymbol{L}\bar{\boldsymbol{\beta}}^*)^T \boldsymbol{P}\dot{\boldsymbol{H}}_i\boldsymbol{P}(\boldsymbol{y}^* - \boldsymbol{L}\bar{\boldsymbol{\beta}}^*) \qquad \text{and} \\
\mathcal{I}(\eta_i, \eta_j) &= \frac{1}{\sigma^2}(\boldsymbol{y} - \boldsymbol{L}\bar{\boldsymbol{\beta}})^T \boldsymbol{P}\dot{\boldsymbol{H}}_i\boldsymbol{T}\dot{\boldsymbol{H}}_j\boldsymbol{P}(\boldsymbol{y} - \boldsymbol{L}\bar{\boldsymbol{\beta}}).
\end{aligned} \tag{8.5.1}$$

Calculating Information

The elements of the information matrix can be efficiently calculated using working variates similar to those in Chapter 3. In particular the working variates for elements corresponding to σ^2 and $\boldsymbol{\eta} = (\boldsymbol{\gamma}^T, \boldsymbol{\theta}^T)$ are

$$\begin{aligned}
\boldsymbol{\omega}_y &= \boldsymbol{y} - \boldsymbol{L}\bar{\boldsymbol{\beta}} \\
\boldsymbol{\omega}_i &= \dot{\boldsymbol{H}}_i\boldsymbol{P}(\boldsymbol{y} - \boldsymbol{L}\bar{\boldsymbol{\beta}}) \\
&= \begin{cases} \boldsymbol{Z}_i\dot{\boldsymbol{G}}_i\boldsymbol{G}_i^{-1}\tilde{\boldsymbol{u}}_i & \text{if } \eta_i = \gamma_i \\ \dot{\boldsymbol{R}}_i\boldsymbol{R}^{-1}\tilde{\boldsymbol{e}} & \text{if } \eta_i = \theta_i \end{cases}
\end{aligned}$$

where $\boldsymbol{Z}_i$ and $\boldsymbol{G}_i$ are the parts of $\boldsymbol{Z}$ and $\boldsymbol{G}$ relating to γ_i, $\dot{\boldsymbol{G}}_i$ is the derivative of $\boldsymbol{G}_i$ with respect to γ_i, $\dot{\boldsymbol{R}}_i$ is the derivative of $\boldsymbol{R}$ with respect to θ_i and $\tilde{\boldsymbol{u}}_i$ and $\tilde{\boldsymbol{e}}$ are the predicted random normal effects corresponding to parameter γ_i and predicted residuals respectively. The quadratic forms are found by appending the matrix $\boldsymbol{\omega} = [\boldsymbol{\omega}_y, \boldsymbol{\omega}_1, \ldots, \boldsymbol{\omega}_m]$ to the coefficient matrix in (8.2.1) in place of $\boldsymbol{y}$. Absorption over the random normal and fixed effect equations gives

$$\begin{aligned}
\boldsymbol{\omega}_y^T\boldsymbol{P}\boldsymbol{\omega}_y &= (\boldsymbol{y} - \boldsymbol{L}\bar{\boldsymbol{\beta}})^T \boldsymbol{P}(\boldsymbol{y} - \boldsymbol{L}\bar{\boldsymbol{\beta}}) \qquad \text{and} \\
\boldsymbol{\omega}_y^T\boldsymbol{P}\boldsymbol{\omega}_i &= (\boldsymbol{y} - \boldsymbol{L}\bar{\boldsymbol{\beta}})^T \boldsymbol{P}\dot{\boldsymbol{H}}_i\boldsymbol{P}(\boldsymbol{y} - \boldsymbol{L}\bar{\boldsymbol{\beta}}).
\end{aligned}$$

Continuing the absorption process over the LASSO equations that have non-zero estimates gives

$$\boldsymbol{\omega}_i^T\boldsymbol{T}\boldsymbol{\omega}_j = (\boldsymbol{y} - \boldsymbol{L}\bar{\boldsymbol{\beta}})^T \boldsymbol{P}\dot{\boldsymbol{H}}_i\boldsymbol{T}\dot{\boldsymbol{H}}_j\boldsymbol{P}(\boldsymbol{y} - \boldsymbol{L}\bar{\boldsymbol{\beta}}).$$

8.5.4 *Solving Adjusted Score Equations*

The adjusted score equations for σ^2 and ϕ have a closed form for the estimates (Theorem 8.6). However, there is no closed form for updating the estimates for $\boldsymbol{\eta}$. Updating is performed by using a method of scoring based on the approximate information in (8.5.1). The approximate information has 0 covariance between ϕ and all other dispersion parameters. This implies that solving the adjusted score equations for $\boldsymbol{\eta}$ can be performed ignoring ϕ. Solving the score equation for $\boldsymbol{\eta}$ also requires iteration. The equation

$$\hat{\boldsymbol{\eta}}^{(i+1)} = \hat{\boldsymbol{\eta}}^{(i)} + \mathcal{I}_{\boldsymbol{\eta}\boldsymbol{\eta}}^{-1} \frac{\partial \ell_r}{\partial \boldsymbol{\eta}} (\hat{\boldsymbol{\eta}}^{(i)})$$

where $\mathcal{I}_{\boldsymbol{\eta}\boldsymbol{\eta}}^{-1}$ is the part of the inverse information corresponding to the random normal dispersion parameters, is repeatedly solved until some predefined convergence is reached.

8.5.5 *Algorithm Overview*

The estimation process proceeds in a parallel manner to the estimation of the dispersion parameters in the model containing only LASSO effects. However, additional steps are required to estimate the extra dispersion parameters ($\boldsymbol{\eta}$). The estimation process is started by specifying an initial set of dispersion parameters. The initial values used as default in the function developed for this thesis are $\eta_i = 0.1$ and the ratio σ^2/ϕ chosen to be 10% of its maximum value. This value for the ratio should allow non-zero LASSO estimates. The initial estimates of σ^2 and ϕ are the estimates obtained from using the unadjusted score equations with q replaced with $q_{\mathcal{S}}$, the number of non-zero LASSO estimates. The initial bootstrap adjustment terms are calculated using these initial dispersion parameter estimates.

Iteration now starts. Estimates for σ^2 and ϕ are updated using the previous estimates and the bootstrap adjustments calculated from the previous estimates. The estimate for $\boldsymbol{\eta}$ is updated according to the average information algorithm outlined in Chapter 3. This step also requires iteration. The bootstrap estimates of the scores are then updated and the iteration process starts again. The estimation process stops when the parameter estimates converge.

The use of the bootstrap routine introduces a stochastic element into the estimation process, and as in Chapter 7 convergence to a point estimate is not guaranteed. This is overcome by using the step halving routine outlined in Chapter 7. Also outlined in Chapter 7 is a method for reducing the number of bootstrap samples needed by initially specifying a low number and then increasing it after each iteration.

8.6 Significance Testing of LASSO Effects

After the dispersion parameters have been estimated, the LASSO estimates are tested for significance using the simulation test of Chapter 7. The null distribution for the LLMM is normal (as all LASSO effects are assumed to be zero), more precisely it is $\mathrm{N}\left(\boldsymbol{X\tau}, \sigma^2 \boldsymbol{H}(\boldsymbol{\eta})\right)$. The parameters $\boldsymbol{\tau}$, σ^2 and $\boldsymbol{\eta}$ are unknown and are replaced by their estimates in the simulation testing procedure.

Both comparison-wise and experiment-wise significant tests are available in a manner completely analogous to those in Chapter 7.

8.7 Simulation

The dispersion parameter estimation method was tested using a small simulation study consisting of simulating and analysing 100 data sets, each comprising $n = 150$ individuals. The simulation model contained: a fixed effect for the mean $\tau_m = 3$, a fixed effect for a covariate whose explanatory variable values were drawn once from a $\text{N}(0, 1)$ distribution and whose effect was $\tau_c = -2$, a set of 15 independent random normal effects with variance $\sigma^2_{15} = 0.6$, a set of 10 independent random normal effects with variance $\sigma^2_{10} = 2$ and a set of LASSO effects with variance $2\phi^2 = 0.1$. The residual variance was set at $\sigma^2 = 1$.

The bootstrap adjustment terms were calculated using $B = 20$ samples initially. This was increased by 10 after each iteration until $B = 200$ was reached. The step size for dispersion parameter updating was decreased to 95% after each iteration. The fixed effect estimates, the dispersion parameter estimates and the number of non-zero LASSO effects were recorded for each analysis.

Results for the simulation are presented in Table 8.1. The fixed effects and dispersion parameters appear to be estimated in an unbiased manner. However, the fixed effect estimates are calculated at the mode, and not the mean, of the joint distribution. This implies that the two coincide within the resolution of this simulation study.

Table 8.1: Results from the efficacy simulation for the LLMM estimated using the bootstrap adjusted scores.

Model Parameter	Simulation Value	Mean	Standard Error
Overall Mean	3.0	2.97	0.086
Covariate	-2.0	-1.99	0.011
Random Effect Variance	0.6	0.61	0.031
Random Effect Variance	2.0	2.15	0.111
LASSO effect Variance	0.1	0.09	0.006
Residual Variance	1.0	1.00	0.013
Non-zero LASSO Estimates	-	12.75	0.218

The variance estimate for the random normal effects with simulated variance of 2.00 may be slightly inflated. However, closer inspection of the results show that there is at least one data set that has a high estimated variance. If this estimate is removed then the mean of the 99 remaining estimates is 2.10 (±0.1). It is unclear if this high value is due to the random data generation process or if it is an artifact of the estimation process.

This simulation shows that the methodology presented in this chapter provides estimates of the fixed effects and the dispersion parameters that are approximately unbiased. The methodology provides the framework where LASSO effects can be jointly modelled with experimental effects.

8.8 Summary

In this chapter the LASSO model was extended to allow for both fixed and random normal effects. Prediction of the model's location effects, given dispersion parameters, was performed using a modified Gaussian elimination technique on an extension of the mixed model equations. The dispersion parameters were estimated by solving adjusted score equations arising from an approximate restricted likelihood. A simulation study showed that this methodology gave approximately unbiased estimates of the fixed effects and the dispersion parameters.

Chapter 9

QTL Detection Using the LLMM

9.1 Introduction

A shortcoming of many QTL mapping methods is the need to fit multiple models to determine the most likely (if any) QTL location(s). Various ways around this problem have been proposed by Whittaker et al. (2000), Gianola et al. (2003) and Xu (2003), who include all the marker information into a single model. However, the number of marker effects in these models is often larger than the size of the mapping population. This means that the marker effects cannot be included into the model as fixed effects, at least not without a model selection step. Even if the number of markers is not greater than the number of individuals a full fixed effect analysis is likely to be ill-conditioned. This may cause unreliable estimates (see Chapter 4 for a discussion of ill-conditioning). To allow fitting all the marker effects into a single model Whittaker et al. (2000), Gianola et al. (2003) and Xu (2003) included the marker effects as random (equivalently constrained) effects. Using the random effects model formulation a natural question to ask is, 'What distribution do the marker effects have?'

Intuitively, it is expected that for any given trait there will be a large number of small QTL effects and a small number of large effects. The distribution of QTL effects, approximately the marker effects, will have high probability near zero with reasonably heavy tails. It is expected that the distribution will also be symmetric as sign is arbitrary for additive effects in most designs. The double exponential distribution may be a reasonable assumption for this distribution. This is supported, at least empirically, by Chamberlin et al. (2005) who investigated the distribution of estimated effects from a series of locations throughout the genome in a dairy cattle population. They showed that the distribution of marker effects has similar distributional moments to that of a double exponential distribution.

Including the marker effects as double exponentially distributed effects forms the LASSO model. In Chapter 7 it was shown that the LASSO random effects model was a powerful tool for estimating effects when there are zero effects in the model. This model was incorporated into a mixed model framework for joint modelling of LASSO effects together with fixed and random normal effects in Chapter 8. This extended mixed model, the LLMM, provides a methodology to map QTL.

In this chapter the details of the method for mapping QTL using the LLMM are presented, assessed by simulation for a trait of large heritability and for a trait of low heritability, and finally demonstrated on the Davies' Gene Mapping data set described in Chapter 1.

9.2 QTL Mapping Method

QTL mapping using the LLMM is a relatively straight-forward procedure. First, a complete set of marker covariates is obtained using the method described in Chapter 2. This procedure replaces the missing and non-informative marker values by the conditional expectations given the flanking markers. For this reason it is important to have an accurate map. However, unlike interval mapping methods, the only use of the marker map in QTL analysis using the LLMM is for the imputation of missing/non-informative markers. This procedure underestimates the variance in the marker covariates. It may be preferable to use procedures which can guard against this, such as multiple imputation. This is not considered in here as it is not considered to be a significant problem.

The scale of the marker variables is changed so that the elements lie within the interval $[-1, 1]$. If a marker was fully informative then the marker variables would have elements of either -1 or 1. Individual marker variables are not standardised to have unit variance as they already have a commensurate scale. This contrasts with the general regression setting (e.g. the prostate data Stamey et al., 1989 analysed in Chapter 7).

The marker effects are included in the LLMM as LASSO effects. They represent QTL surrogates as they will be correlated to the true QTL through genetic linkage. Other genetic effects and experimental effects are included in the model as either fixed or random normal effects, depending on what is considered appropriate.

The dispersion parameters and the fixed, random normal and the LASSO effects are estimated using the methods described in Chapter 8. The LASSO effects are tested using an experiment-wise (called a genome-wise) test at a pre-specified significance level using the simulation method in Chapter 8. Those marker effects that are significant are defined to be associated with a QTL. That is, there is a QTL near these markers. Many of the marker effects will be estimated to be zero, which should make model interpretation easier by screening out sections of the genome without QTL.

This method of identifying and estimating important marker effects assumes that the effects are random. Implicitly, this assumes that the probability of an individual marker effect being identically zero is precisely zero. From a genetical viewpoint this is unrealistic and is therefore a potential flaw in the model. This flaw could be guarded against by using the LLMM as a model selection step to identify the marker effects that are non-zero and then by estimating their effects using an unbiased estimator. Unfortunately, this style of estimation introduces another flaw, as such estimates are likely to be biased upwards (e.g. Broman, 2001). This is an example of the statistical phenomenon called ‘selection bias’ (Miller, 2002, Chapter 6). Ideally, estimation of effect size should come from a secondary study. However, this is generally not practical and estimation must come from the same data as does identification.

If selection bias is ignored then unbiased estimates for the marker effects can still be obtained using the LLMM QTL mapping framework. The LLMM could be used as a *working* model, used solely to identify the important marker effects. Estimation of marker effects would then proceed by treating them as fixed effects in the corresponding (normal) mixed model. The estimated variance components from the LLMM should provide excellent starting values for this mixed model and hence computation can be minimal. This two step procedure parallels the general model selection method outlined in Öjelund et al. (2001) and Efron et al. (2004). The basic idea behind these published methods is to use the LASSO to select the model and then to estimate the non-zero effects via least squares. Unfortunately, neither sets of authors provide simulation results to see if the two-step working model method is superior to the simpler LASSO model. That is, to see if selection bias detracts substantially from the intuitive appeal of the two-step estimation method.

Due to the potential selection bias, the two-step estimation method is not used in this thesis. Also, the primary focus of the methods in this chapter is the identification of QTL and not the estimation of their effects. Where the size of an effect is presented, it is based on the random effects model.

9.3 Simulations

The simulation design was based on that in Broman & Speed (2002). Briefly, genomes of 9 chromosomes of 100cM in length were simulated. Each chromosome contained 11 equally spaced, fully informative markers. Chromosomes 1 and 2 each contained 2 QTL located at 30cM and 70cM, the QTL were linked in coupling phase on chromosome 1 and in repulsive phase on chromosome 2. Chromosomes 3, 4 and 5 each contained 1 QTL at positions 50cM, 30cM and 0cM respectively. The remaining chromosomes did not contain any QTL. Four population sizes were considered $n = 130$, $n = 250$ and $n = 500$. The population size $n = 130$ was chosen to mirror a single family from the Davies' Gene Mapping data. The others are used as a comparison to the methods presented in Broman & Speed (2002). For each population size the simulation study consisted of generating and analysing 500 data sets.

The LLMM method was compared against the recommended approach from Broman & Speed (2002), namely the forward selection method with entry criterion defined by BIC$_\delta$ (FS-BIC$_\delta$). The BIC$_\delta$ was created by multiplying BIC by a constant $\delta \geq 1$, chosen to account for multiple testing (Broman & Speed, 2002). For this simulation design $\delta = 2.43$ for $n = 130$, $\delta = 2.10$ for $n = 250$ and $\delta = 1.85$ for $n = 500$. Any marker variable included in the final model was considered to be linked to a QTL.

Only FS-BIC$_\delta$ was used for comparison as the relative merit of FS-BIC$_\delta$ against many other analyses was presented in Broman & Speed (2002). These authors found that FS-BIC$_\delta$ was amongst the best of methods considered at identifying QTL. In particular, the FS-BIC$_\delta$ method behaved favourably when compared to interval mapping and composite interval mapping (Chapter 2).

The LLMM method was implemented using a convergence tolerance of 0.002, chosen to be as small as possible while still maintaining computational feasibility. The number of bootstrap samples for the first iteration was 30, this was increased by 5 each iteration until the number of samples reached 100. The step size was reduced after each iteration to 95% of its previous value. Only those effects that were significant at this level were considered to be associated with a QTL.

Two criteria were used to classify if a QTL was correctly identified or not. Firstly, a QTL was considered to be correctly identified if its associated marker was found to be significant (0cM tolerance). Secondly, a QTL was considered to be correctly identified if its associated marker or one either side of it on the linkage map was chosen as being significant (10cM tolerance). The 10cM tolerance criterion is the one used by Broman & Speed (2002). For some simulated data sets more than 1 marker within 10cM of a true QTL was identified as significant. Under the 10cM tolerance criterion these situations were deemed to find the QTL only once.

Two types of traits were considered. Those with a high heritability ($h^2 = 50\%$) and those with a low heritability ($h^2 = 10\%$). The QTL substitution effect (replacing one QTL allele with the other) for the highly heritable trait is 0.760 and for the lowly heritable trait was 0.252. The highly heritable trait was considered in Broman & Speed (2002).

9.3.1 *Heritability 50%*

For all population sizes and tolerance levels the average number of correctly identified QTL are given in Table 9.1. Results are given for correctly chosen total number of QTL, linked QTL (in both phases) and unlinked QTL. For nearly all the population sizes and both tolerance levels the LLMM identified more QTL than the FS-BIC$_\delta$ method. The exceptions were for $n = 250$ (tolerance 10cM) and $n = 500$ (tolerance 10cM). The differences are small in both these cases. The difference in identification rates for the two tolerance levels was less for the LLMM than for FS-BIC$_\delta$, implying the LLMM was more precise in locating QTL.

The average number of extraneous QTL identified are presented in Table 9.2. Results are given for the total extraneous QTL, the extraneous QTL linked to a true QTL and the extraneous QTL not linked to a true QTL. For the smaller population sizes ($n = 130$ and $n = 250$), the LLMM identified approximately equal numbers of extraneous QTL for tolerance 10cM and less for tolerance 0cM. For population size $n = 500$, the LLMM might identify slightly more extraneous QTL than FS-BIC$_\delta$. However, for tolerance 0cM most of the difference arises from multiple significant marker effects within 10cM of the true QTL. In practice these would imply a single genetic effect within the region.

Table 9.1: Heritability 50% QTL simulation results: the average number of correctly identified QTL per simulated data set.

			Correctly Chosen QTL			
Tolerance	#Individuals	Model	Total	Coupling	Repulsion	Other
0cM	n=130	LLMM	3.61 (0.06)	1.13 (0.03)	0.70 (0.03)	1.78 (0.04)
		FS-BIC$_\delta$	2.68 (0.07)	0.83 (0.03)	0.37 (0.03)	1.48 (0.04)
	n=250	LLMM	6.03 (0.04)	1.75 (0.02)	1.61 (0.03)	2.67 (0.02)
		FS-BIC$_\delta$	5.68 (0.05)	1.58 (0.03)	1.48 (0.03)	2.62 (0.03)
	n=500	LLMM	6.93 (0.01)	1.98 (0.01)	1.97 (0.01)	2.98 (0.01)
		FS-BIC$_\delta$	6.80 (0.02)	1.93 (0.01)	1.94 (0.01)	2.93 (0.01)
10cM	n=130	LLMM	4.22 (0.06)	1.37 (0.03)	0.81 (0.03)	2.04 (0.04)
		FS-BIC$_\delta$	3.40 (0.07)	1.09 (0.02)	0.48 (0.03)	1.84 (0.04)
	n=250	LLMM	6.33 (0.04)	1.87 (0.02)	1.68 (0.03)	2.78 (0.02)
		FS-BIC$_\delta$	6.42 (0.04)	1.84 (0.02)	1.69 (0.03)	2.89 (0.02)
	n=500	LLMM	6.99 (0.01)	2.00 (0.00)	1.99 (0.00)	3.00 (0.00)
		FS-BIC$_\delta$	7.00 (0.00)	2.00 (0.00)	2.00 (0.00)	3.00 (0.00)

9.3.2 *Heritability 10%*

For all population sizes and tolerance levels the average number of correctly identified QTL are given in Table 9.3. Results are given for correctly chosen total number of QTL, linked QTL (in both phases) and unlinked QTL. The reduction in heritability reduces the ability of all methods to identify QTL. However, the FS-BIC$_\delta$ method identifies more QTL for the smallest population size and approximately equivalent numbers for the larger population sizes. The difference between the number of correctly identified QTL under the two tolerance levels is greater for FS-BIC$_\delta$. This implies that the QTL identified by the LLMM are more precise in location than those identified by FS-BIC$_\delta$.

The average number of extraneous QTL identified are presented in Table 9.4. Results are given for the total extraneous QTL, the extraneous QTL linked to a true QTL and the extraneous QTL not linked to a true QTL. The number of extraneous QTL identified is substantially less for the LLMM than for FS-BIC$_\delta$. This is particularly true for the small population sizes.

9.3.3 *Simulation Summary*

The LLMM appears to be a useful tool for identifying QTL when compared to FS-BIC$_\delta$ which was shown to be competitive compared to some other commonly used methods (Broman & Speed, 2002).

For large heritabilities, the LLMM identified more QTL than FS-BIC$_\delta$ without increasing the rate of false positives. For low heritabilities, the LLMM identified less QTL but had a

Table 9.2: Heritability 50% QTL simulation results: the average number of extraneous QTL found per simulated data set.

			Extraneous QTL		
Tolerance	#Individuals	Model	Total	Linked	Unlinked
0cM	n=130	LLMM	1.12 (0.05)	1.04 (0.03)	0.08 (0.01)
		FS-BIC$_\delta$	1.16 (0.04)	1.11 (0.04)	0.05 (0.01)
	n=250	LLMM	0.86 (0.04)	0.83 (0.04)	0.03 (0.01)
		FS-BIC$_\delta$	1.11 (0.04)	1.08 (0.04)	0.03 (0.01)
	n=500	LLMM	0.58 (0.03)	0.52 (0.03)	0.05 (0.01)
		FS-BIC$_\delta$	0.44 (0.03)	0.41 (0.03)	0.03 (0.01)
10cM	n=130	LLMM	0.41 (0.03)	0.33 (0.03)	0.08 (0.01)
		FS-BIC$_\delta$	0.43 (0.03)	0.38 (0.03)	0.05 (0.01)
	n=250	LLMM	0.29 (0.02)	0.26 (0.02)	0.03 (0.01)
		FS-BIC$_\delta$	0.32 (0.02)	0.29 (0.02)	0.03 (0.01)
	n=500	LLMM	0.18 (0.02)	0.13 (0.02)	0.05 (0.01)
		FS-BIC$_\delta$	0.11 (0.02)	0.08 (0.01)	0.03 (0.01)

substantially lower rate of false positives. The LLMM may be conservative but greater confidence can be placed in a QTL identified by the LLMM (as opposed to one identified by FS-BIC$_\delta$). Also, the location of the QTL identified by the LLMM appeared to be more accurate than the FS-BIC$_\delta$ method.

9.4 Davies' Gene Mapping Data

The genetic model used for the analysis of the Davies' Gene Mapping experiment (Chapter 1) included a separate (LASSO) effect for each informative marker within each of the three sires. This model allowed the effect of a QTL to differ between sires, both in location and in size. It was likely that some QTL would not be segregating in all sires and this model allowed for that. There were a total of 570 unique marker within sire effects included as LASSO effects in the model. The model contained only additive non-epistatic effects.

Missing and non-informative marker values were assigned the conditional expectation given the flanking markers. The cattle linkage map from Ihara et al. (2004) was used for this purpose. This linkage map does not contain locations for single nucleotide polymorphisms (SNPs) which were located on chromosomes BTA1, BTA8, BTA11 and BTA20. The SNPs were inserted into the linkage map using the CRI-MAP software (Green et al., 1990). This software assesses a given putative location for a SNP by calculating the likelihood from a multi-point linkage analysis (Lander & Green, 1987). After the linkage map was generated, the QTL Express software (Seaton et al., 2002) was used to calculate the conditional probabilities at specified locations along the linkage map. The marker locations form a subset of

Table 9.3: Heritability 10% QTL simulation results: the average number of correctly identified QTL per simulated data set.

			Correctly Chosen QTL			
Tolerance	#Individuals	Model	Total	Coupling	Repulsion	Other
0cM	n=130	LLMM	0.17 (0.02)	0.10 (0.01)	0.01 (0.00)	0.07 (0.01)
		FS-BIC$_\delta$	0.35 (0.02)	0.17 (0.02)	0.03 (0.01)	0.15 (0.02)
	n=250	LLMM	0.51 (0.03)	0.28 (0.02)	0.02 (0.01)	0.21 (0.02)
		FS-BIC$_\delta$	0.57 (0.03)	0.33 (0.02)	0.02 (0.01)	0.23 (0.02)
	n=500	LLMM	1.33 (0.04)	0.58 (0.27)	0.11 (0.02)	0.63 (0.03)
		FS-BIC$_\delta$	1.27 (0.04)	0.55 (0.23)	0.09 (0.01)	0.63 (0.03)
10cM	n=130	LLMM	0.29 (0.02)	0.17 (0.02)	0.02 (0.01)	0.11 (0.02)
		FS-BIC$_\delta$	0.64 (0.02)	0.35 (0.02)	0.05 (0.01)	0.24 (0.02)
	n=250	LLMM	0.74 (0.03)	0.42 (0.02)	0.03 (0.01)	0.29 (0.02)
		FS-BIC$_\delta$	0.95 (0.03)	0.56 (0.02)	0.04 (0.01)	0.34 (0.02)
	n=500	LLMM	1.71 (0.05)	0.78 (0.03)	0.14 (0.02)	0.77 (0.03)
		FS-BIC$_\delta$	1.76 (0.04)	0.82 (0.02)	0.12 (0.02)	0.82 (0.04)

the locations from QTL Express.

The model contained fixed effects for the overall mean, breed of dam factor and date of birth covariate. The breed of dam factor was coded using a dummy variable. The model also contained a single set of 6 random normal effects for Cohort (sex by year interaction). This means that the model contained 3 fixed effects and 3 dispersion parameters to be estimated. There were 6 random normal effects and 570 LASSO effects to be predicted.

The estimation process used a convergence tolerance of 0.001. The number of bootstrap samples used for calculating the expected scores was initially 500, which was increased by 20 after each iteration until a maximum of 1000 was reached. This may be excessive, but the analysis was considered to be important enough to warrant these choices. After each iteration the step size for dispersion parameter updating was reduced to 97.5% of its previous value. The initial values of the residual variance and the variance associated with the Cohort factor were chosen to be those from a corresponding mixed normal model without the marker effects. The initial value for the dispersion parameter ϕ was taken to be 0.2. Reasonable starting values and a large number of bootstrap samples meant that the initial step length (see Section 8.5.4) could be reduced as the initial parameter values should be close to the final parameter estimates. The initial step length was taken to be one half of the step chosen by the scoring algorithm. After estimation of the dispersion parameters, the marker effect predictions were tested using a genome-wise simulation test at the 5% type-1 error rate. The empirical null-distribution was found using 2000 simulated data sets.

The fixed effect and dispersion parameter estimates are given in Table 9.5. There were

Table 9.4: Heritability 10% QTL simulation results: the average number of extraneous QTL found per simulated data set.

			Extraneous QTL		
Tolerance	#Individuals	Model	Total	Linked	Unlinked
0cM	n=130	LLMM	0.27 (0.03)	0.22 (0.02)	0.05 (0.01)
		FS-BIC$_\delta$	0.71 (0.02)	0.54 (0.02)	0.16 (0.02)
	n=250	LLMM	0.44 (0.03)	0.41 (0.03)	0.03 (0.01)
		FS-BIC$_\delta$	0.66 (0.03)	0.61 (0.03)	0.05 (0.01)
	n=500	LLMM	0.67 (0.03)	0.64 (0.03)	0.03 (0.01)
		FS-BIC$_\delta$	0.79 (0.03)	0.76 (0.03)	0.03 (0.01)
10cM	n=130	LLMM	0.14 (0.02)	0.09 (0.01)	0.05 (0.01)
		FS-BIC$_\delta$	0.41 (0.02)	0.25 (0.02)	0.16 (0.02)
	n=250	LLMM	0.20 (0.02)	0.18 (0.02)	0.03 (0.01)
		FS-BIC$_\delta$	0.29 (0.02)	0.24 (0.02)	0.05 (0.01)
	n=500	LLMM	0.27 (0.02)	0.24 (0.02)	0.03 (0.01)
		FS-BIC$_\delta$	0.30 (0.02)	0.28 (0.02)	0.03 (0.01)

only 5 non-zero marker estimates, presented in Table 9.6. The small number of non-zero estimates implies that there are large regions of the genome that do not contain any QTL for birth weight. Using the LLMM model these regions are easily identified. The two non-zero marker estimates on BTA1 for sire 368 are the two markers defining a single interval. This result would imply the presence of a QTL within that interval.

Previous analysis of these data (Morris et al., 2003) used an interval mapping procedure (see Chapter 2) within each sire family. Also, they used the theoretical genome-wise threshold obtained from the method of Lander & Kruglyak (1995). They identified QTL on chromosomes BTA1, BTA5, BTA14 and BTA21 in roughly the same locations and sire families to those non-zero estimates presented in Table 9.6. The simulation results presented in this chapter and in Broman & Speed (2002) suggest that the significant results in Morris et al. (2003) may be due to false positives from either the interval mapping method or from the determination of the critical threshold value.

This analysis shows the flexibility and power of the QTL mapping using the LLMM. In particular, all the genetic effects are included in a single model which contains effects for the sensible modelling of non-QTL genetic and experimental sources of variation. Also, interpretation of the marker (LASSO) effects is a relatively easy task as most of them have been predicted to be zero.

Table 9.5: Fixed effect and dispersion parameter estimates for the birth weight trait from the Davies' Gene Mapping Data. Units of outcome variable are kilograms (kilograms per day for day of birth covariate).

	Parameter	Estimate
Fixed Effects	μ	16.569
	Breed of Dam (Jersey)	0.000
	Breed of Dam (Limousin)	8.799
	Sire (LOU)	0.000
	Sire (RYAN)	1.539
	Sire (TOM)	0.037
	Day of Birth	0.048
Dispersion Parameters	var (Cohort)	2.49
	var (Marker Effects)	0.02
	Residual Variance	10.37

Table 9.6: Davies' Gene Mapping Data analysis of birth weight trait. The non-zero marker estimates (units are kilograms). Estimates are the substitution effect of replacing one sire marker allele with the other.

Chromosome	Position (cM)	Sire	Substitution Effect
BTA1	99.2	368	0.043
BTA1	109.9	368	0.132
BTA5	35.3	398	0.478
BTA14	50.9	361	0.855*
BTA21	12.6	368	0.348

* significant at the 5% genome-wise level (critical value 0.66), all others not significant.

9.5 Summary

In this chapter, the performance of the LLMM for identifying QTL was assessed via simulation and demonstrated using the Davies' Gene Mapping data. The simulations showed that the LLMM was competitive for identifying QTL. The analysis of the Davies' Gene Mapping data for the birth weight trait identified only five non-zero marker effects. This clearly shows which regions of the genome do not contain QTL and enables localisation of potential QTL.

Chapter 10

An Alternative LASSO Random Effects Model

10.1 Introduction and Motivation

The LASSO was investigated from a random effects model viewpoint in Chapter 7. Specifically, the LASSO effects were assumed to come from a double exponential distribution. The dispersion parameters were estimated by maximising an approximate likelihood for the observed data. The resulting score equations were biased and were adjusted using an estimate of their respective expected values. This estimate was obtained empirically using a computationally expensive bootstrap routine. A simulation showed that this approach is a superior, or at least competitive, method for obtaining the LASSO predictions and possesses the desirable attribute of arising from a statistical model.

A better approach would be to derive an alternative approximate likelihood whose score equations were not biased and could be solved analytically, without a stochastic element to the maximisation. Estimation using these alternative score equations could reduce computation drastically. In this chapter an alternative LASSO random model is explored. It leads to an approximate likelihood that produces dispersion parameter estimates which have a tolerable level of bias and that is computationally simpler to maximise.

An unsatisfactory feature of the estimation method presented in Chapter 7 is that the dispersion parameter ϕ is estimated to be zero when the LASSO effects are all predicted to be zero (see Theorem 7.4). The alternative method presented in this chapter does not possess this attribute.

Up to this point in this thesis, prediction of the LASSO random effects has used the standard LASSO method proposed initially by Tibshirani (1996). This predictor is the mode of the predictive (posterior) distribution $f(\boldsymbol{\beta}|\boldsymbol{y})$. This is not the only type of predictor available. In particular, best prediction and best linear prediction (e.g. Searle et al., 1992) can be evaluated for the LASSO random effects model. The alternative form of the LASSO random effects model also suggests that prediction conditional on the random variance effects could be considered. These different types of predictors are compared for predictive performance using a simulation study.

Generally, in a mixed model framework the distribution $f(\boldsymbol{\beta}|\boldsymbol{y})$ is called the 'predictive distribution' as $f(\boldsymbol{\beta}|\boldsymbol{y}) = f(\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}})$, where $\tilde{\boldsymbol{\beta}}$ is the best predictor of $\boldsymbol{\beta}$ (Searle et al., 1992). This terminology is used throughout this chapter, not the Bayesian term 'posterior distribution'.

In Chapter 7, a method of inference was proposed. It tested the null hypothesis that $\beta_i = 0$ against the alternative that $\beta_i \neq 0$. This essentially treats the LASSO random effects as though they were fixed. Formal inference for random effects should be based on the predictive distribution, in particular probability statements such as $\mathrm{P}(\beta_i > 0|\boldsymbol{y})$ should be used. Using these probabilities, inference about the random effect can be formally obtained. Two methods for calculating these probabilities are proposed.

S-PLUS code that performs the methods described in this chapter is presented in Appendix D.

10.2 The Alternative Random Effects Model

The alternative random LASSO model stems from the fact that there is a conditional normal equivalent to the double exponential distribution (Andrews & Mallows, 1974 and Kotz et al., 2001). Yuan & Lin (2005) formulate the central double exponential distribution using the hierarchical structure

$$\beta|\delta \sim \mathrm{N}(0, \delta) \qquad \text{and} \qquad \delta \sim \exp(2\phi^2)$$

where $\exp(2\phi^2)$ is the exponential distribution with mean $2\phi^2$. The density function of δ is

$$f(\delta) = \frac{1}{2\phi^2}\exp\left(-\frac{1}{2\phi^2}\delta\right).$$

This alternative representation of the double exponential distribution has interesting implications for the LASSO model. In particular, it shows that there are equivalences between the LASSO and a variant of ridge regression called 'adaptive ridge regression'. This is not a new result (Grandvalet, 1998 and Grandvalet & Canu, 1998) but the alternative random effects model formulation formalises the equivalence on a probabilistic level rather than a purely numerical level.

The LASSO random effects model can be stated using the alternative formulation as

$$\boldsymbol{y} = \boldsymbol{L\beta} + \boldsymbol{e} \tag{10.2.1}$$

where

$$\begin{aligned} \beta_i|\delta_i &\sim \mathrm{N}(0, \delta_i); & i = 1 \ldots q \\ \delta_i &\sim \exp(2\phi^2) & \text{and} \\ \boldsymbol{e} &\sim \mathrm{N}(\boldsymbol{0}, \sigma^2\boldsymbol{I}). & \end{aligned}$$

The model still only has two parameters (σ^2 and $2\phi^2$) but now there are twice as many random effects ($\{\beta_i\}$ and $\{\delta_i\}$). The advantage of the alternative model to that presented in Chapter 7 is that the joint distribution of the observed data $\boldsymbol{y}$, the LASSO effects $\boldsymbol{\beta}$ and the vector of random variance effects $\boldsymbol{\delta}$ is differentiable with respect to $\boldsymbol{\beta}$ and $\boldsymbol{\delta}$. This implies that Laplace's approximation can be used directly without resorting to analogues of derivatives from convex programming. This is likely to have been the cause of the biased estimates of the methods presented in Chapter 7.

10.3 Approximate Marginal Likelihood

The likelihood is based on the marginal distribution of the observed data. Using the alternative random effects model the marginal distribution is found by integrating the joint distribution $f(\boldsymbol{y}, \boldsymbol{\beta}, \boldsymbol{\delta})$ over both $\boldsymbol{\beta}$ and $\boldsymbol{\delta}$. Fortunately, the joint distribution $f(\boldsymbol{y}, \boldsymbol{\beta}|\boldsymbol{\delta})$ of the observed data and the LASSO effects is normal and hence the marginal distribution of $\boldsymbol{y}$ over $\boldsymbol{\beta}$ conditional on $\boldsymbol{\delta}$ is also normal. In particular,

$$\boldsymbol{y}|\boldsymbol{\delta} \sim \mathrm{N}(\mathbf{0}, \boldsymbol{H})$$

where $\boldsymbol{H} = \sigma^2 \boldsymbol{I} + \boldsymbol{L}\boldsymbol{D}\boldsymbol{L}^T$, $\boldsymbol{D} = \mathrm{diagv}(\boldsymbol{\delta})$ and $\mathrm{diagv}(\boldsymbol{x})$ is a square matrix with the elements of $\boldsymbol{x}$ filling its diagonals.

An exact analytical expressions for the marginal likelihood is difficult to obtain. A Laplace approximation (Chapter 3) to the marginal likelihood is used in its place. The random variables to be integrated out are defined on the positive set of real numbers, not the full set. It is expected that a transformation of the random effects will make the approximation more accurate. In the following Theorem the transformation $\boldsymbol{\alpha} = \log(\boldsymbol{\delta})$ is used.

Theorem 10.1. *The approximate marginal log-likelihood for the alternative LASSO random effects model is*

$$\ell(\sigma^2, 2\phi^2; \boldsymbol{y}) = -q\log\left(2\phi^2\right) - \frac{1}{2}\log|\bar{\boldsymbol{\Lambda}}_o| - \frac{1}{2}\log|\bar{\boldsymbol{H}}| \\ - \frac{1}{2}\boldsymbol{y}^T\bar{\boldsymbol{H}}^{-1}\boldsymbol{y} - \frac{1}{2\phi^2}\sum_{i=1}^{q}\exp(\bar{\alpha}_i) + \sum_{i=1}^{q}\bar{\alpha}_i$$

where $\bar{\boldsymbol{\alpha}}$ is the maximiser of

$$Q(\boldsymbol{\alpha}; \boldsymbol{y}, \sigma^2, 2\phi^2) = -\frac{1}{2}\log|\boldsymbol{H}| - \frac{1}{2}\boldsymbol{y}^T\boldsymbol{H}^{-1}\boldsymbol{y} - \frac{1}{2\phi^2}\sum_{i=1}^{q}\exp(\alpha_i) + \sum_{i=1}^{q}\alpha_i,$$

$\bar{\boldsymbol{H}}$ is $\boldsymbol{H}$ evaluated at $\boldsymbol{\delta} = \exp(\bar{\boldsymbol{\alpha}})$ and $\bar{\boldsymbol{\Lambda}}_o$ is

$$\boldsymbol{\Lambda}_o = \frac{\partial^T Q(\boldsymbol{\alpha})}{\partial\boldsymbol{\alpha}\partial\boldsymbol{\alpha}^T} = -\mathrm{diagv}\left(\exp(\boldsymbol{\alpha}) \circ \boldsymbol{b}\right) + \\ \exp(\boldsymbol{\alpha})\exp(\boldsymbol{\alpha})^T \circ \left(\frac{1}{2}\left(\boldsymbol{L}^T\boldsymbol{H}^{-1}\boldsymbol{L}\right) \circ \left(\boldsymbol{L}^T\boldsymbol{H}^{-1}\boldsymbol{L}\right) - \boldsymbol{w}\boldsymbol{w}^T \circ \left(\boldsymbol{L}^T\boldsymbol{H}^{-1}\boldsymbol{L}\right)\right)$$

evaluated at $\boldsymbol{\alpha} = \bar{\boldsymbol{\alpha}}$.

Proof. The exact marginal distribution is

$$\begin{aligned} f(\boldsymbol{y}) &= \int_{\mathbb{R}^q_+} f(\boldsymbol{y}|\boldsymbol{\delta})f(\boldsymbol{\delta})\partial\boldsymbol{\delta} \\ &= \frac{1}{(2\pi)^{\frac{n}{2}}(2\phi^2)^q}\int_{\mathbb{R}^q_+}|\boldsymbol{H}|^{-\frac{1}{2}}\exp\left(-\frac{1}{2}\boldsymbol{y}^T\boldsymbol{H}^{-1}\boldsymbol{y}-\frac{1}{2\phi^2}\sum_{i=1}^q\delta_i\right)\partial\boldsymbol{\delta} \\ &= \frac{1}{(2\pi)^{\frac{n}{2}}(2\phi^2)^q}\int_{\mathbb{R}^q_+}\exp\left(-\frac{1}{2}\log|\boldsymbol{H}|-\frac{1}{2}\boldsymbol{y}^T\boldsymbol{H}^{-1}\boldsymbol{y}-\frac{1}{2\phi^2}\sum_{i=1}^q\delta_i\right)\partial\boldsymbol{\delta} \end{aligned}$$

where $\mathbb{R}^q_+$ is the positive orthant (inclusive of zero) of $\mathbb{R}^q$. Using the transformation $\boldsymbol{\alpha} = \log(\boldsymbol{\delta})$ the marginal distribution is

$$\begin{aligned} f(\boldsymbol{y}) &= \frac{1}{(2\pi)^{\frac{n}{2}}(2\phi^2)^q}\int_{\mathbb{R}^q}\exp\left(-\frac{1}{2}\log|\boldsymbol{H}|-\frac{1}{2}\boldsymbol{y}^T\boldsymbol{H}^{-1}\boldsymbol{y}-\frac{1}{2\phi^2}\sum_{i=1}^q\exp(\alpha_i)\right)\left|\frac{\partial\boldsymbol{\delta}}{\partial\boldsymbol{\alpha}}\right|\partial\boldsymbol{\alpha} \\ &= \frac{1}{(2\pi)^{\frac{n}{2}}(2\phi^2)^q}\int_{\mathbb{R}^q}\exp\left(-\frac{1}{2}\log|\boldsymbol{H}|-\frac{1}{2}\boldsymbol{y}^T\boldsymbol{H}^{-1}\boldsymbol{y}-\frac{1}{2\phi^2}\sum_{i=1}^q\exp(\alpha_i)+\log|\boldsymbol{D}|\right)\partial\boldsymbol{\alpha} \\ &= \frac{1}{(2\pi)^{\frac{n}{2}}(2\phi^2)^q}\int_{\mathbb{R}^q}\exp\left(-\frac{1}{2}\log|\boldsymbol{H}|-\frac{1}{2}\boldsymbol{y}^T\boldsymbol{H}^{-1}\boldsymbol{y}-\frac{1}{2\phi^2}\sum_{i=1}^q\exp(\alpha_i)+\sum_{i=1}^q\alpha_i\right)\partial\boldsymbol{\alpha} \\ &= \frac{1}{(2\pi)^{\frac{n}{2}}(2\phi^2)^q}\int_{\mathbb{R}^q}\exp(Q(\boldsymbol{\alpha}))\,\partial\boldsymbol{\alpha} \qquad \text{say} \end{aligned}$$

where $\boldsymbol{H}$ and $\boldsymbol{D}$ are now parameterised in terms of $\boldsymbol{\alpha}$ (not $\boldsymbol{\delta}$), and $|\partial\boldsymbol{\delta}/\partial\boldsymbol{\alpha}|$ is the Jacobian of the transformation.

The Laplace approximation requires the first two derivatives of $Q(\boldsymbol{\alpha})$ with respect to $\boldsymbol{\alpha}$. Consider initially the derivatives with respect to a single α_i.

$$\begin{aligned} \frac{\partial Q(\boldsymbol{\alpha})}{\partial\alpha_i} &= -\frac{1}{2}\text{tr}\left(\boldsymbol{H}^{-1}\frac{\partial\boldsymbol{H}}{\partial\alpha_i}\right)+\frac{1}{2}\boldsymbol{y}^T\boldsymbol{H}^{-1}\frac{\partial\boldsymbol{H}}{\partial\alpha_i}\boldsymbol{H}^{-1}\boldsymbol{y}-\frac{1}{2\phi^2}\exp(\alpha_i)+1 \\ &= -\frac{1}{2}\exp(\alpha_i)\,\boldsymbol{l}_i^T\boldsymbol{H}^{-1}\boldsymbol{l}_i+\frac{1}{2}\exp(\alpha_i)\,\boldsymbol{y}^T\boldsymbol{H}^{-1}\boldsymbol{l}_i\boldsymbol{l}_i^T\boldsymbol{H}^{-1}\boldsymbol{y}-\frac{1}{2\phi^2}\exp(\alpha_i)+1 \\ &= -\frac{1}{2}\exp(\alpha_i)\,M_{ii}+\frac{1}{2}\exp(\alpha_i)\,w_i^2-\frac{1}{2\phi^2}\exp(\alpha_i)+1 \\ &= -\exp(\alpha_i)\left(\frac{1}{2}M_{ii}-\frac{1}{2}w_i^2+\frac{1}{2\phi^2}\right)+1 \\ &= -\exp(\alpha_i)\,b_i+1 \qquad \text{say} \end{aligned}$$

where $\boldsymbol{l}_i$ is the i^{th} column of the design matrix $\boldsymbol{L}$, M_{ij} is the $(i,j)^{th}$ element of $\boldsymbol{M} = \boldsymbol{L}^T\boldsymbol{H}^{-1}\boldsymbol{L}$ and $w_i = \boldsymbol{l}_i^T\boldsymbol{H}^{-1}\boldsymbol{y}$. Also,

$$\begin{aligned} \frac{\partial^2 Q(\boldsymbol{\alpha})}{\partial\alpha_i^2} &= -\exp(\alpha_i)\,b_i-\exp(\alpha_i)\left(-\frac{1}{2}\exp(\alpha_i)\,\boldsymbol{l}_i^T\boldsymbol{H}^{-1}\boldsymbol{l}_i\boldsymbol{l}_i^T\boldsymbol{H}^{-1}\boldsymbol{l}_i+\exp(\alpha_i)\,w_i\boldsymbol{l}_i^T\boldsymbol{H}^{-1}\boldsymbol{l}_iw_i\right) \\ &= -\exp(\alpha_i)\,b_i+\exp(\alpha_i)^2\left(\frac{1}{2}M_{ii}^2-w_i^2M_{ii}\right) \end{aligned}$$

and for $i \neq j$

$$\begin{aligned}\frac{\partial^2 Q(\boldsymbol{\alpha})}{\partial\alpha_i\partial\alpha_j} &= -\exp(\alpha_i)\left(-\frac{1}{2}\exp(\alpha_j)\,\boldsymbol{l}_i^T\boldsymbol{H}^{-1}\boldsymbol{l}_j\boldsymbol{l}_j^T\boldsymbol{H}^{-1}\boldsymbol{l}_i + \exp(\alpha_j)\,w_i\boldsymbol{l}_i^T\boldsymbol{H}^{-1}\boldsymbol{l}_j\boldsymbol{l}_j^T\boldsymbol{H}^{-1}\boldsymbol{y}\right)\\ &= \exp(\alpha_i)\exp(\alpha_j)\left(\frac{1}{2}M_{ij}M_{ij} - w_iw_jM_{ij}\right).\end{aligned}$$

The first and second derivatives with respect to the vector $\boldsymbol{\alpha}$ are therefore

$$\begin{aligned}\frac{\partial Q(\boldsymbol{\alpha})}{\partial\boldsymbol{\alpha}} &= -\exp(\boldsymbol{\alpha})\circ\left(\frac{1}{2}\text{vecd}\left(\boldsymbol{L}^T\boldsymbol{H}^{-1}\boldsymbol{L}\right) - \frac{1}{2}\boldsymbol{w}\circ\boldsymbol{w} + \frac{1}{2\phi^2}\boldsymbol{1}\right) + \boldsymbol{1}\\ &= -\exp(\boldsymbol{\alpha})\circ\boldsymbol{b} + \boldsymbol{1} \qquad\qquad \text{and} \qquad\qquad (10.3.1)\\ \frac{\partial^2 Q(\boldsymbol{\alpha})}{\partial\boldsymbol{\alpha}\partial\boldsymbol{\alpha}^T} &= -\text{diagv}\left(\exp(\boldsymbol{\alpha})\circ\boldsymbol{b}\right) +\\ &\qquad \exp(\boldsymbol{\alpha})\exp(\boldsymbol{\alpha})^T\circ\left(\frac{1}{2}\left(\boldsymbol{L}^T\boldsymbol{H}^{-1}\boldsymbol{L}\right)\circ\left(\boldsymbol{L}^T\boldsymbol{H}^{-1}\boldsymbol{L}\right) - \boldsymbol{w}\boldsymbol{w}^T\circ\left(\boldsymbol{L}^T\boldsymbol{H}^{-1}\boldsymbol{L}\right)\right)\\ &= -\boldsymbol{\Lambda}_o \qquad\qquad \text{say,} \qquad\qquad (10.3.2)\end{aligned}$$

where $\boldsymbol{X}\circ\boldsymbol{Z}$ is the element by element multiplication of the same sized matrices (or vectors) $\boldsymbol{X}$ and $\boldsymbol{Z}$, $\text{vecd}\left(\boldsymbol{X}\right)$ is a vector consisting of the diagonal elements of the square matrix $\boldsymbol{X}$ and $\boldsymbol{1}$ is a vector of ones.

The approximation to the marginal distribution is

$$\begin{aligned}f(\boldsymbol{y}) &= \frac{1}{(2\pi)^{\frac{n}{2}}(2\phi^2)^q}\int_{\mathbb{R}^q}\exp\left(Q(\boldsymbol{\alpha})\right)\partial\boldsymbol{\alpha}\\ &\approx \frac{1}{(2\pi)^{\frac{n}{2}}(2\phi^2)^q}(2\pi)^{\frac{q}{2}}|\bar{\boldsymbol{\Lambda}}_o|^{-\frac{1}{2}}\exp\left(Q(\bar{\boldsymbol{\alpha}})\right)\int_{\mathbb{R}^q}\frac{1}{(2\pi)^{\frac{q}{2}}}|\bar{\boldsymbol{\Lambda}}_o|^{\frac{1}{2}}\exp\left(-\frac{1}{2}(\boldsymbol{\alpha}-\bar{\boldsymbol{\alpha}})^T\bar{\boldsymbol{\Lambda}}_o(\boldsymbol{\alpha}-\bar{\boldsymbol{\alpha}})\right)\partial\boldsymbol{\alpha}\\ &= \frac{1}{(2\pi)^{\frac{(n-q)}{2}}(2\phi^2)^q}|\bar{\boldsymbol{\Lambda}}_o|^{-\frac{1}{2}}\exp\left(Q(\bar{\boldsymbol{\alpha}})\right)\end{aligned}$$

where $\bar{\boldsymbol{\Lambda}}_o$ and $\bar{\boldsymbol{H}}$ are $\boldsymbol{\Lambda}_o$ and $\boldsymbol{H}$ respectively evaluated at $\boldsymbol{\alpha} = \bar{\boldsymbol{\alpha}}$.

The log-likelihood is obtained by taking the log of the marginal distribution and disregarding terms not involving the parameters σ^2 and $2\phi^2$. □

The matrix $\boldsymbol{\Lambda}_o$ is the observed information matrix for $\boldsymbol{\alpha}$. It is not guaranteed to be positive definite. In some situations it may be convenient to use the expected information conditional on the random variance effects, which is guaranteed to be positive definite.

Theorem 10.2. *The conditional expected information for $\boldsymbol{\alpha}$ is*

$$\boldsymbol{\Lambda}_e = \text{diagv}\left(\frac{1}{2\phi^2}\exp(\boldsymbol{\alpha})\right) + \frac{1}{2}\exp(\boldsymbol{\alpha})\exp(\boldsymbol{\alpha})^T\circ\left(\boldsymbol{L}^T\boldsymbol{H}^{-1}\boldsymbol{L}\right)\circ\left(\boldsymbol{L}^T\boldsymbol{H}^{-1}\boldsymbol{L}\right).$$

Proof. Define $\boldsymbol{w}$ and $\boldsymbol{b}$ as in the proof of Theorem 10.1. Then

$$\begin{aligned}\text{E}\left(w_iw_j|\boldsymbol{\alpha}\right) &= \text{cov}\left(w_i, w_j|\boldsymbol{\alpha}\right) + \text{E}\left(w_i|\boldsymbol{\alpha}\right)\text{E}\left(w_j|\boldsymbol{\alpha}\right)\\ &= \boldsymbol{l}_i^T\boldsymbol{H}^{-1}\text{var}\left(\boldsymbol{y}|\boldsymbol{\alpha}\right)\boldsymbol{H}^{-1}\boldsymbol{l}_j\\ &= \boldsymbol{l}_i\boldsymbol{H}^{-1}\boldsymbol{l}_j \qquad\qquad \text{and}\\ \text{E}\left(\boldsymbol{w}\boldsymbol{w}^T|\boldsymbol{\alpha}\right) &= \boldsymbol{L}^T\boldsymbol{H}^{-1}\boldsymbol{L}.\end{aligned}$$

The conditional expected information for $\boldsymbol{\alpha}$ is

$$\begin{aligned}
\mathrm{E}\left(\boldsymbol{\Lambda}_o|\boldsymbol{\alpha}\right) &= \operatorname{diagv}\left(\exp\left(\boldsymbol{\alpha}\right) \circ \mathrm{E}\left(\boldsymbol{b}|\boldsymbol{\alpha}\right)\right) + \frac{1}{2}\exp\left(\boldsymbol{\alpha}\right)\exp\left(\boldsymbol{\alpha}\right)^T \circ \left(\left(\boldsymbol{L}^T\boldsymbol{H}^{-1}\boldsymbol{L}\right) \circ \left(\boldsymbol{L}^T\boldsymbol{H}^{-1}\boldsymbol{L}\right)\right) \\
&= \operatorname{diagv}\left(\frac{1}{2\phi^2}\exp\left(\boldsymbol{\alpha}\right)\right) + \frac{1}{2}\exp\left(\boldsymbol{\alpha}\right)\exp\left(\boldsymbol{\alpha}\right)^T \circ \left(\boldsymbol{L}^T\boldsymbol{H}^{-1}\boldsymbol{L}\right) \circ \left(\boldsymbol{L}^T\boldsymbol{H}^{-1}\boldsymbol{L}\right).
\end{aligned}$$

□

Both observed and conditional expected information matrices have been used with success in Laplace approximations. The justification for using the conditional expected information comes from an appeal to the so-called Bayesian Central Limit Theorem. This theorem states that as the sample size increases $f(\boldsymbol{y}|\boldsymbol{\alpha})$ will tend to dominate over $f(\boldsymbol{\alpha})$ in the joint distribution. Hence, the information matrix of the conditional distribution will tend to the information matrix of the joint distribution. Also, the conditional distribution will tend to be normal. In this application of Laplace's method the choice of information matrix that will give the most stable and accurate approximation is not immediately clear and derivation will continue with both. Unless the distinction between $\boldsymbol{\Lambda}_o$ and $\boldsymbol{\Lambda}_e$ is needed, the generic notation of $\boldsymbol{\Lambda}$ will be used to mean either information matrix. Score equations and estimate attributes will be derived for both.

10.3.1 *Estimating the Random Dispersion Effects*

The estimates $\bar{\boldsymbol{\alpha}}$ (the maximisers of the integrand of the joint distribution) are required to evaluate the log-likelihood in Theorem 10.1. This is performed using the method of scoring with the expected information for $\boldsymbol{\alpha}$ in Theorem 10.2. The update step for the $(i+1)^{th}$ iteration is

$$\bar{\boldsymbol{\alpha}}^{(i+1)} = \bar{\boldsymbol{\alpha}}^{(i)} + s\boldsymbol{\Lambda}_e^{-1}\frac{\partial Q(\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}}$$

where both $\boldsymbol{\Lambda}_e$ and $\partial Q(\boldsymbol{\alpha})/\partial\boldsymbol{\alpha}$ are evaluated at $\boldsymbol{\alpha} = \bar{\boldsymbol{\alpha}}^{(i)}$ and $0 < s \leq 1$. In the implementation of this procedure, the step size s was chosen to be 1 for all iterations except the first where $s = 1/50$. This initial step size prevents absurd movements in the first iteration. If no sensible starting values are available for $\bar{\boldsymbol{\alpha}}$, then $\alpha_i = \log\left(2\phi^2\right) - 1/2$ is used for all i. This is an approximate expected value for α_i as

$$\begin{aligned}
\mathrm{E}\left(\alpha_i\right) &= \mathrm{E}\left(\log\delta_i\right) \\
&\approx \mathrm{E}\left(\log 2\phi^2 + (\delta_i - 2\phi^2)\frac{1}{2\phi^2} - \frac{1}{2}(\delta_i - 2\phi^2)^2\frac{1}{4\phi^4}\right) \\
&= \log 2\phi^2 - \frac{1}{8\phi^4}\operatorname{var}\left(\delta_i\right) \\
&= \log 2\phi^2 - \frac{1}{2}.
\end{aligned}$$

Not surprisingly, sensible starting values, if available, speed up convergence substantially.

The iterative procedure is terminated when both the scores are zero within a predefined tolerance and successive iterations' estimates ($\bar{\boldsymbol{\alpha}}$) do not change within that same predefined tolerance. The default tolerance in this implementation is 10^{-6}. This may seem excessive but subsequent numerical methods require accurate values of $\bar{\boldsymbol{\alpha}}$.

The maximisation process appears to be stable and sufficiently fast for all the modelling situations considered in this chapter. The computational cost increases with the number of effects but particularly with the sample size, due to the calculation of $\boldsymbol{H}^{-1}$.

10.3.2 *An Example of the Approximate Likelihoods' Surfaces*

A single realisation of the approximate likelihood surfaces (as a function of σ^2 and $2\phi^2$) may give an indication of the behaviour of the likelihood in general. The approximate likelihood is evaluated with both the observed information and the conditional expected information for a simulated data set. The data were generated using the random effects model in (10.2.1) with $n = 50$, $q = 20$, $\sigma^2 = 1$, $2\phi^2 = 1$ and the elements of $\boldsymbol{L}$ were generated from a standard normal. The resulting likelihoods are presented in Figure 10.1.

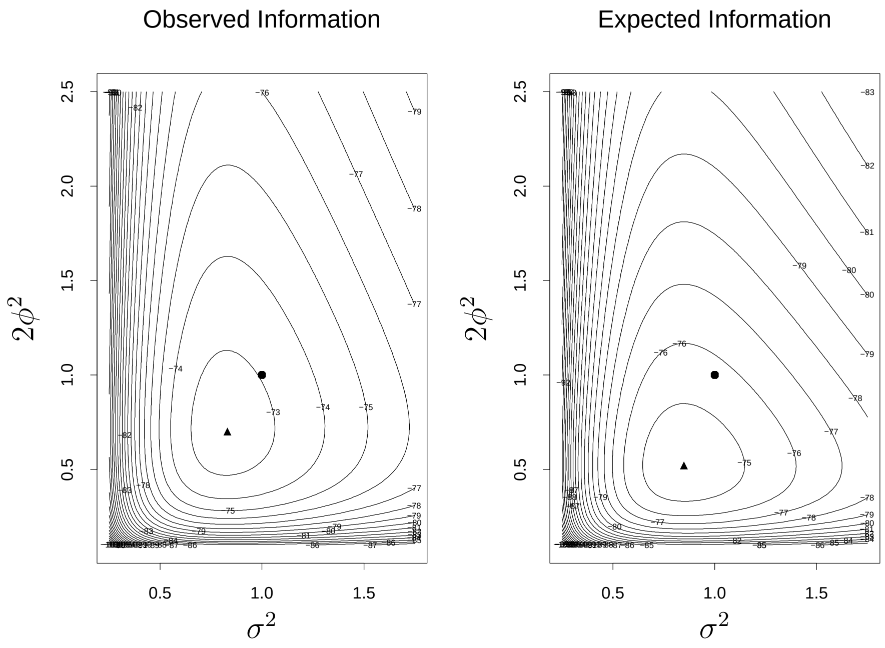

Figure 10.1: An example of the approximate likelihoods given in Theorem 10.1 using the observed and conditional expected information matrices for $\boldsymbol{\alpha}$. Contours are at every 1.0 unit change of the likelihood. The circular points (●) are the values used for simulation and the triangular points (▲) are the observed maxima.

Both approximate likelihoods have a distinct maximum and appear to be concave. This implies that the task of finding the maximum can be reduced to finding a local maximum. The difference between the two approximations is quite large. Most notably the maxima are in quite different locations and the rate of decrease as either σ^2 or $2\phi^2$ increase is greater for the likelihood based on the expected information. The difference in location for the maxima suggests that one or both of the approximations could be biased. This is investigated using a simulation study later in the chapter.

10.4 Score Equations and Their Components

The score equations require, amongst other relatively easily obtainable expressions, analytical expressions for the derivatives of $\bar{\boldsymbol{\alpha}}$ and $\bar{\boldsymbol{\Lambda}}$.

Theorem 10.3. *The derivatives of* $\bar{\boldsymbol{\alpha}}$ *are*

$$\frac{\partial\bar{\boldsymbol{\alpha}}}{\partial\sigma^2} = \left(\left(\frac{1}{2}\bar{\boldsymbol{M}}\circ\bar{\boldsymbol{M}} - \bar{\boldsymbol{w}}\bar{\boldsymbol{w}}^T\circ\bar{\boldsymbol{M}}\right)\bar{\boldsymbol{D}} - \bar{\boldsymbol{D}}^{-1}\right)^{-1} \times \left(\bar{\boldsymbol{w}}\circ\bar{\boldsymbol{w}}_2 - \frac{1}{2}\text{vecd}\left(\bar{\boldsymbol{M}}_2\right)\right)$$

and

$$\frac{\partial\bar{\boldsymbol{\alpha}}}{\partial 2\phi^2} = -\frac{1}{4\phi^4}\left(\left(\frac{1}{2}\bar{\boldsymbol{M}}\circ\bar{\boldsymbol{M}} - \bar{\boldsymbol{w}}\bar{\boldsymbol{w}}^T\circ\bar{\boldsymbol{M}}\right)\bar{\boldsymbol{D}} - \bar{\boldsymbol{D}}^{-1}\right)^{-1}\boldsymbol{1}$$

where $\bar{\boldsymbol{D}} = \text{diagv}\left(\exp\left(\bar{\boldsymbol{\alpha}}\right)\right)$, $\bar{\boldsymbol{w}}_2 = \boldsymbol{L}^T\bar{\boldsymbol{H}}^{-2}\boldsymbol{y}$, $\bar{\boldsymbol{M}}_2 = \boldsymbol{L}^T\bar{\boldsymbol{H}}^{-2}\boldsymbol{L}$ *and* $\bar{\boldsymbol{H}}^{-2} = \bar{\boldsymbol{H}}^{-1}\bar{\boldsymbol{H}}^{-1}$.

Proof. There is no closed form for $\bar{\boldsymbol{\alpha}}$. However, the derivatives can be found using implicit differentiation of the estimating equation for $\boldsymbol{\alpha}$ in (10.3.1). Consider the derivative with respect to σ^2, $\partial Q(\boldsymbol{\alpha})/\partial\alpha_i = 0$ implies that

$$0 = -\frac{1}{2}\boldsymbol{l}_i^T\bar{\boldsymbol{H}}^{-1}\boldsymbol{l}_i + \frac{1}{2}\boldsymbol{y}^T\bar{\boldsymbol{H}}^{-1}\boldsymbol{l}_i\boldsymbol{l}_i^T\bar{\boldsymbol{H}}^{-1}\boldsymbol{y} - \frac{1}{2\phi^2} + \exp\left(-\bar{\alpha}_i\right)$$

therefore

$$0 = \frac{1}{2}\boldsymbol{l}_i^T\bar{\boldsymbol{H}}^{-1}\frac{\partial\bar{\boldsymbol{H}}}{\partial\sigma^2}\bar{\boldsymbol{H}}^{-1}\boldsymbol{l}_i - \boldsymbol{y}^T\bar{\boldsymbol{H}}^{-1}\boldsymbol{l}_i\boldsymbol{l}_i^T\bar{\boldsymbol{H}}^{-1}\frac{\partial\bar{\boldsymbol{H}}}{\partial\sigma^2}\bar{\boldsymbol{H}}^{-1}\boldsymbol{y} - \exp\left(-\bar{\alpha}_i\right)\frac{\partial\bar{\alpha}_i}{\partial\sigma^2}.$$

Now

$$\begin{aligned}\frac{\partial\bar{\boldsymbol{H}}}{\partial\sigma^2} &= \boldsymbol{I} + \boldsymbol{L}\text{diagv}(\exp\bar{\boldsymbol{\alpha}}\circ\frac{\partial\bar{\boldsymbol{\alpha}}}{\partial\sigma^2})\boldsymbol{L}^T\\ &= \boldsymbol{I} + \boldsymbol{L}\bar{\boldsymbol{D}}_{\sigma^2}\boldsymbol{L}^T \qquad\qquad \text{say.}\end{aligned}$$

So

$$\begin{aligned}&\bar{w}_i\boldsymbol{l}_i^T\bar{\boldsymbol{H}}^{-2}\boldsymbol{y} - \frac{1}{2}\boldsymbol{l}_i^T\bar{\boldsymbol{H}}^{-2}\boldsymbol{l}_i\\ &\quad= \frac{1}{2}\boldsymbol{l}_i^T\bar{\boldsymbol{H}}^{-1}\boldsymbol{L}\bar{\boldsymbol{D}}_{\sigma^2}\boldsymbol{L}^T\bar{\boldsymbol{H}}^{-1}\boldsymbol{l}_i - \bar{w}_i\boldsymbol{l}_i^T\bar{\boldsymbol{H}}^{-1}\boldsymbol{L}\bar{\boldsymbol{D}}_{\sigma^2}\boldsymbol{L}^T\bar{\boldsymbol{H}}^{-1}\boldsymbol{y} - \exp\left(-\bar{\alpha}_i\right)\frac{\partial\bar{\alpha}_i}{\partial\sigma^2}\\ &\quad= \sum_{j=1}^{q}\left(\frac{1}{2}\bar{M}_{ij}^2\exp\left(\bar{\alpha}_j\right)\frac{\partial\bar{\alpha}_j}{\partial\sigma^2}\right) - \bar{w}_i\sum_{j=1}^{q}\left(\bar{M}_{ij}\bar{w}_j\exp\left(\bar{\alpha}_j\right)\frac{\partial\bar{\alpha}_j}{\partial\sigma^2}\right) - \exp\left(-\bar{\alpha}_i\right)\frac{\partial\bar{\alpha}_i}{\partial\sigma^2}\\ &\quad= \sum_{j=1}^{q}\left(\left(\frac{1}{2}\bar{M}_{ij}^2 - \bar{w}_i\bar{w}_j\bar{M}_{ij}\right)\exp\left(\bar{\alpha}_j\right)\frac{\partial\bar{\alpha}_j}{\partial\sigma^2}\right) - \exp\left(-\bar{\alpha}_i\right)\frac{\partial\bar{\alpha}_i}{\partial\sigma^2}\\ &\quad= \left(\frac{1}{2}\bar{M}_{i\cdot}\circ\bar{M}_{i\cdot} - \bar{w}_i\bar{\boldsymbol{w}}^T\circ\bar{M}_{i\cdot}\right)\left(\exp\left(\bar{\boldsymbol{\alpha}}\right)\circ\frac{\partial\bar{\boldsymbol{\alpha}}}{\partial\sigma^2}\right) - \exp\left(-\bar{\alpha}_i\right)\frac{\partial\bar{\alpha}_i}{\partial\sigma^2}.\end{aligned}$$

For all the α_i

$$\left(\left(\frac{1}{2}\bar{\boldsymbol{M}} \circ \bar{\boldsymbol{M}} - \bar{\boldsymbol{w}}\bar{\boldsymbol{w}}^T \circ \bar{\boldsymbol{M}}\right) \bar{\boldsymbol{D}} - \operatorname{diagv}\left(\exp\left(-\bar{\boldsymbol{\alpha}}\right)\right)\right) \frac{\partial \bar{\boldsymbol{\alpha}}}{\partial \sigma^2} = \\ \bar{\boldsymbol{w}} \circ \left(\boldsymbol{L}^T \bar{\boldsymbol{H}}^{-2} \boldsymbol{y}\right) - \frac{1}{2}\operatorname{vecd}\left(\boldsymbol{L}^T \bar{\boldsymbol{H}}^{-2} \boldsymbol{L}\right).$$

A similar derivation gives the derivative with respect to $2\phi^2$. □

Theorem 10.4. *The derivatives of the observed information matrix for $\boldsymbol{\alpha}$ evaluated at $\bar{\boldsymbol{\alpha}}$ are*

$$\begin{aligned}
\frac{\partial \bar{\boldsymbol{\Lambda}}_o}{\partial \sigma^2} &= \bar{\boldsymbol{D}}_{\sigma^2} \operatorname{diagv}\left(\frac{1}{2}\operatorname{vecd}\left(\bar{\boldsymbol{M}}\right) - \frac{1}{2}\bar{\boldsymbol{w}} \circ \bar{\boldsymbol{w}} + \frac{1}{2\phi^2}\mathbf{1}\right) \\
&+ \operatorname{diagv}\left(\exp\left(\bar{\boldsymbol{\alpha}}\right) \circ \left(-\frac{1}{2}\operatorname{vecd}\left(\bar{\boldsymbol{M}}_2\right) - \frac{1}{2}\operatorname{vecd}\left(\bar{\boldsymbol{M}}\bar{\boldsymbol{D}}_{\sigma^2}\bar{\boldsymbol{M}}\right) + \bar{\boldsymbol{w}} \circ \bar{\boldsymbol{w}}_2 + \bar{\boldsymbol{w}} \circ \bar{\boldsymbol{M}}\bar{\boldsymbol{D}}_{\sigma^2}\bar{\boldsymbol{w}}\right)\right) \\
&- \left(\bar{\boldsymbol{\Delta}}_{\sigma^2} + \bar{\boldsymbol{\Delta}}_{\sigma^2}^T\right) \circ \left(\frac{1}{2}\bar{\boldsymbol{M}} \circ \bar{\boldsymbol{M}} - \bar{\boldsymbol{w}}\bar{\boldsymbol{w}}^T \circ \bar{\boldsymbol{M}}\right) \\
&- \exp\left(\bar{\boldsymbol{\alpha}}\right)\exp\left(\bar{\boldsymbol{\alpha}}\right)^T \circ \left(-\bar{\boldsymbol{M}} \circ \bar{\boldsymbol{M}}_2 + \left(\bar{\boldsymbol{w}}\bar{\boldsymbol{w}}_2^T + \bar{\boldsymbol{w}}_2\bar{\boldsymbol{w}}^T\right) \circ \bar{\boldsymbol{M}} + \bar{\boldsymbol{w}}\bar{\boldsymbol{w}}^T \circ \bar{\boldsymbol{M}}_2\right) \\
&- \exp\left(\bar{\boldsymbol{\alpha}}\right)\exp\left(\bar{\boldsymbol{\alpha}}\right)^T \circ \left(\left(\bar{\boldsymbol{w}}\bar{\boldsymbol{w}}^T - \bar{\boldsymbol{M}}\right) \circ \bar{\boldsymbol{M}}\bar{\boldsymbol{D}}_{\sigma^2}\bar{\boldsymbol{M}} + \left(\bar{\boldsymbol{w}}\left(\bar{\boldsymbol{M}}\bar{\boldsymbol{D}}_{\sigma^2}\bar{\boldsymbol{w}}\right)^T + \left(\bar{\boldsymbol{M}}\bar{\boldsymbol{D}}_{\sigma^2}\bar{\boldsymbol{w}}\right)\bar{\boldsymbol{w}}^T\right) \circ \bar{\boldsymbol{M}}\right)
\end{aligned}$$

and

$$\begin{aligned}
\frac{\partial \bar{\boldsymbol{\Lambda}}_o}{\partial 2\phi^2} &= \bar{\boldsymbol{D}}_{2\phi^2} \operatorname{diagv}\left(\frac{1}{2}\operatorname{vecd}\left(\bar{\boldsymbol{M}}\right) - \frac{1}{2}\bar{\boldsymbol{w}}\bar{\boldsymbol{w}}^T + \frac{1}{2\phi^2}\mathbf{1}\right) \\
&+ \operatorname{diagv}\left(\exp\left(\bar{\boldsymbol{\alpha}}\right) \circ \left(-\frac{1}{2}\operatorname{vecd}\left(\bar{\boldsymbol{M}}\bar{\boldsymbol{D}}_{2\phi^2}\bar{\boldsymbol{M}}\right) + \bar{\boldsymbol{w}} \circ \bar{\boldsymbol{M}}\bar{\boldsymbol{D}}_{2\phi^2}\bar{\boldsymbol{w}} - \frac{1}{4\phi^4}\mathbf{1}\right)\right) \\
&- \left(\bar{\boldsymbol{\Delta}}_{2\phi^2} + \bar{\boldsymbol{\Delta}}_{2\phi^2}^T\right) \circ \left(\frac{1}{2}\bar{\boldsymbol{M}} \circ \bar{\boldsymbol{M}} - \bar{\boldsymbol{w}}\bar{\boldsymbol{w}}^T \circ \bar{\boldsymbol{M}}\right) \\
&- \exp\left(\bar{\boldsymbol{\alpha}}\right)\exp\left(\bar{\boldsymbol{\alpha}}\right)^T \circ \left(\left(\bar{\boldsymbol{w}}\bar{\boldsymbol{w}}^T - \bar{\boldsymbol{M}}\right) \circ \bar{\boldsymbol{M}}\bar{\boldsymbol{D}}_{2\phi^2}\bar{\boldsymbol{M}} + \left(\bar{\boldsymbol{w}}\left(\bar{\boldsymbol{M}}\bar{\boldsymbol{D}}_{2\phi^2}\bar{\boldsymbol{w}}\right)^T + \left(\bar{\boldsymbol{M}}\bar{\boldsymbol{D}}_{2\phi^2}\bar{\boldsymbol{w}}\right)\bar{\boldsymbol{w}}^T\right) \circ \bar{\boldsymbol{M}}\right)
\end{aligned}$$

where $\bar{\boldsymbol{D}}_{\sigma^2} = \operatorname{diagv}\left(\exp\left(\boldsymbol{\alpha}\right) \circ \frac{\partial \bar{\boldsymbol{\alpha}}}{\partial \sigma^2}\right)$, $\bar{\boldsymbol{\Delta}}_{\sigma^2} = (\exp\left(\bar{\boldsymbol{\alpha}}\right) \circ \partial \bar{\boldsymbol{\alpha}}/\partial \sigma^2)\exp\left(\bar{\boldsymbol{\alpha}}\right)^T$, *and* $\bar{\boldsymbol{D}}_{2\phi^2}$ *and* $\bar{\boldsymbol{\Delta}}_{2\phi^2}$ *are similarly defined.*

Proof. Consider the derivative with respect to σ^2. The derivative with respect to $2\phi^2$ follows similarly. Express the observed information for $\boldsymbol{\alpha}$, given in (10.3.2), as

$$\bar{\boldsymbol{\Lambda}}_o = \boldsymbol{A} + \boldsymbol{B}.$$

The derivatives of $\bar{\boldsymbol{M}}$ and $\bar{\boldsymbol{w}}$ repeatedly occur throughout the derivation. They are

$$\frac{\partial \bar{M}_{ij}}{\partial \sigma^2} = -\boldsymbol{l}_i^T \bar{\boldsymbol{H}}^{-2} \boldsymbol{l}_j - \boldsymbol{l}_i^T \bar{\boldsymbol{H}}^{-1} \boldsymbol{L} \bar{\boldsymbol{D}}_{\sigma^2} \boldsymbol{L}^T \bar{\boldsymbol{H}}^{-1} \boldsymbol{l}_j \qquad \text{and}$$

$$\frac{\partial \bar{w}_i}{\partial \sigma^2} = -\boldsymbol{l}_i \bar{\boldsymbol{H}}^{-2} \boldsymbol{y} - \boldsymbol{l}_i^T \bar{\boldsymbol{H}}^{-1} \boldsymbol{L} \bar{\boldsymbol{D}}_{\sigma^2} \boldsymbol{L}^T \bar{\boldsymbol{H}}^{-1} \boldsymbol{y}.$$

The matrix $\boldsymbol{A}$ is diagonal and the derivative of the i^{th} diagonal element A_{ii} is

$$\begin{aligned}\frac{\partial A_{ii}}{\partial \sigma^2} &= \exp\left(\bar{\alpha}_i\right) \frac{\partial \bar{\alpha}_i}{\partial \sigma^2} \left(\frac{1}{2}\bar{M}_{ii} - \frac{1}{2}\bar{w}_i^2 + \frac{1}{2\phi^2}\right) \\ &\quad + \exp\left(\bar{\alpha}_i\right) \left(-\frac{1}{2}\boldsymbol{l}_i^T \bar{\boldsymbol{H}}^{-2} \boldsymbol{l}_i - \frac{1}{2}\boldsymbol{l}_i^T \bar{\boldsymbol{H}}^{-1} \boldsymbol{L} \bar{\boldsymbol{D}}_{\sigma^2} \boldsymbol{L}^T \bar{\boldsymbol{H}}^{-1} \boldsymbol{l}_i + \bar{w}_i \boldsymbol{l}_i \bar{\boldsymbol{H}}^{-2} \boldsymbol{y} + \bar{w}_i \boldsymbol{l}_i^T \bar{\boldsymbol{H}}^{-1} \boldsymbol{L} \bar{\boldsymbol{D}}_{\sigma^2} \bar{\boldsymbol{w}}\right).\end{aligned}$$

The derivative of the full matrix $\boldsymbol{A}$ is

$$\begin{aligned}\frac{\partial \boldsymbol{A}}{\partial \sigma^2} &= \bar{\boldsymbol{D}}_{\sigma^2} \mathrm{diagv}\left(\frac{1}{2}\mathrm{vecd}\left(\bar{\boldsymbol{M}}\right) - \frac{1}{2}\bar{\boldsymbol{w}} \circ \bar{\boldsymbol{w}} + \frac{1}{2\phi^2}\mathbf{1}\right) \\ &\quad + \mathrm{diagv}\left(\exp\left(\bar{\boldsymbol{\alpha}}\right) \circ \left(-\frac{1}{2}\mathrm{vecd}\left(\bar{\boldsymbol{M}}_2\right) - \frac{1}{2}\mathrm{vecd}\left(\bar{\boldsymbol{M}} \bar{\boldsymbol{D}}_{\sigma^2} \bar{\boldsymbol{M}}\right) + \bar{\boldsymbol{w}} \circ \bar{\boldsymbol{w}}_2 + \bar{\boldsymbol{w}} \circ \bar{\boldsymbol{M}} \bar{\boldsymbol{D}}_{\sigma^2} \bar{\boldsymbol{w}}\right)\right).\end{aligned}$$

The derivative of the $(i,j)^{th}$ element of $\boldsymbol{B}$, B_{ij} say, is

$$\begin{aligned}\frac{\partial B_{ij}}{\partial \sigma^2} &= -\left(\exp\left(\bar{\alpha}_i\right) \frac{\partial \bar{\alpha}_i}{\partial \sigma^2} \exp\left(\bar{\alpha}_j\right) + \exp\left(\bar{\alpha}_i\right) \exp\left(\bar{\alpha}_j\right) \frac{\partial \bar{\alpha}_j}{\partial \sigma^2}\right) \left(\frac{1}{2}\bar{M}_{ij}^2 - \bar{w}_i \bar{w}_j \bar{M}_{ij}\right) \\ &\quad - \exp\left(\bar{\alpha}_i\right) \exp\left(\bar{\alpha}_j\right) \left(-\bar{M}_{ij} \boldsymbol{l}_i^T \bar{\boldsymbol{H}}^{-2} \boldsymbol{l}_j + \bar{w}_j \boldsymbol{l}_i^T \bar{\boldsymbol{H}}^{-2} \boldsymbol{y} \bar{M}_{ij} + \bar{w}_i \boldsymbol{l}_j^T \bar{\boldsymbol{H}}^{-2} \boldsymbol{y} \bar{M}_{ij} + \bar{w}_i \bar{w}_j \boldsymbol{l}_i^T \bar{\boldsymbol{H}}^{-2} \boldsymbol{l}_j\right) \\ &\qquad - \exp\left(\bar{\alpha}_i\right) \exp\left(\bar{\alpha}_j\right) \Big(-\bar{M}_{ij} \boldsymbol{l}_i^T \bar{\boldsymbol{H}}^{-1} \boldsymbol{L} \bar{\boldsymbol{D}}_{\sigma^2} \boldsymbol{L}^T \bar{\boldsymbol{H}}^{-1} \boldsymbol{l}_j + \bar{w}_j \boldsymbol{l}_i^T \bar{\boldsymbol{H}}^{-1} \boldsymbol{L} \bar{\boldsymbol{D}}_{\sigma^2} \boldsymbol{L}^T \bar{\boldsymbol{H}}^{-1} \boldsymbol{y} \bar{M}_{ij} \\ &\qquad\qquad + \bar{w}_i \boldsymbol{l}_j^T \bar{\boldsymbol{H}}^{-1} \boldsymbol{L} \bar{\boldsymbol{D}}_{\sigma^2} \boldsymbol{L}^T \bar{\boldsymbol{H}}^{-1} \boldsymbol{y} \bar{M}_{ij} + \bar{w}_i \bar{w}_j \boldsymbol{l}_i^T \bar{\boldsymbol{H}}^{-1} \boldsymbol{L} \bar{\boldsymbol{D}}_{\sigma^2} \boldsymbol{L}^T \bar{\boldsymbol{H}}^{-1} \boldsymbol{l}_j\Big)\end{aligned}$$

and the full matrix derivative is

$$\begin{aligned}\frac{\partial \boldsymbol{B}}{\partial \sigma^2} &= -\left(\bar{\boldsymbol{\Delta}}_{\sigma^2} + \bar{\boldsymbol{\Delta}}_{\sigma^2}^T\right) \circ \left(\frac{1}{2}\bar{\boldsymbol{M}} \circ \bar{\boldsymbol{M}} - \bar{\boldsymbol{w}} \bar{\boldsymbol{w}}^T \circ \bar{\boldsymbol{M}}\right) \\ &\qquad - \exp\left(\bar{\boldsymbol{\alpha}}\right) \exp\left(\bar{\boldsymbol{\alpha}}\right)^T \circ \left(-\bar{\boldsymbol{M}} \circ \bar{\boldsymbol{M}}_2 + \left(\bar{\boldsymbol{w}} \bar{\boldsymbol{w}}_2^T + \bar{\boldsymbol{w}}_2 \bar{\boldsymbol{w}}^T\right) \circ \bar{\boldsymbol{M}} + \bar{\boldsymbol{w}} \bar{\boldsymbol{w}}^T \circ \bar{\boldsymbol{M}}_2\right) \\ &- \exp\left(\bar{\boldsymbol{\alpha}}\right) \exp\left(\bar{\boldsymbol{\alpha}}\right)^T \circ \left(\left(\bar{\boldsymbol{w}} \bar{\boldsymbol{w}}^T - \bar{\boldsymbol{M}}\right) \circ \bar{\boldsymbol{M}} \bar{\boldsymbol{D}}_{\sigma^2} \bar{\boldsymbol{M}} + \left(\bar{\boldsymbol{w}} \left(\bar{\boldsymbol{M}} \bar{\boldsymbol{D}}_{\sigma^2} \bar{\boldsymbol{w}}\right)^T + \bar{\boldsymbol{M}} \bar{\boldsymbol{D}}_{\sigma^2} \bar{\boldsymbol{w}} \bar{\boldsymbol{w}}^T\right) \circ \bar{\boldsymbol{M}}\right).\end{aligned}$$

The derivative of the estimated observed information $\bar{\boldsymbol{\Lambda}}_o$ is $\partial \boldsymbol{A}/\partial \sigma^2 + \partial \boldsymbol{B}/\partial \sigma^2$. □

Theorem 10.5. *The derivatives for the conditional expected information for $\boldsymbol{\alpha}$ evaluated at $\bar{\boldsymbol{\alpha}}$ are*

$$\frac{\partial \bar{\boldsymbol{\Lambda}}_e}{\partial \sigma^2} = \frac{1}{2\phi^2}\bar{\boldsymbol{D}}_{\sigma^2} + \frac{1}{2}\left(\bar{\boldsymbol{\Delta}}_{\sigma^2} + \bar{\boldsymbol{\Delta}}_{\sigma^2}^T\right) \circ \bar{\boldsymbol{M}} \circ \bar{\boldsymbol{M}} - \exp\left(\bar{\boldsymbol{\alpha}}\right) \exp\left(\bar{\boldsymbol{\alpha}}\right)^T \circ \bar{\boldsymbol{M}} \circ \left(\bar{\boldsymbol{M}}_2 + \bar{\boldsymbol{M}} \bar{\boldsymbol{D}}_{\sigma^2} \bar{\boldsymbol{M}}\right)$$

and

$$\frac{\partial \bar{\boldsymbol{\Lambda}}_e}{\partial 2\phi^2} = -\frac{1}{4\phi^4}\bar{\boldsymbol{D}} + \frac{1}{2\phi^2}\bar{\boldsymbol{D}}_{2\phi^2} + \frac{1}{2}\left(\bar{\boldsymbol{\Delta}}_{2\phi^2} + \bar{\boldsymbol{\Delta}}_{2\phi^2}^T\right) \circ \bar{\boldsymbol{M}} \circ \bar{\boldsymbol{M}} - \exp\left(\bar{\boldsymbol{\alpha}}\right) \exp\left(\bar{\boldsymbol{\alpha}}\right)^T \circ \bar{\boldsymbol{M}} \circ \left(\bar{\boldsymbol{M}} \bar{\boldsymbol{D}}_{2\phi^2} \bar{\boldsymbol{M}}\right).$$

Proof. Consider the derivative with respect to σ^2. The derivative with respect to $2\phi^2$ follows similarly. Express the conditional expected information matrix in Theorem (10.2) as

$$\bar{\boldsymbol{\Lambda}}_e = \boldsymbol{A} + \boldsymbol{B}.$$

The terms $\boldsymbol{A}$ and $\boldsymbol{B}$ will be treated separately.

$$\frac{\partial \boldsymbol{A}}{\partial \sigma^2} = \frac{1}{2\phi^2}\bar{\boldsymbol{D}} \circ \operatorname{diagv}\left(\frac{\partial \bar{\boldsymbol{\alpha}}}{\partial \sigma^2}\right) = \frac{1}{2\phi^2}\bar{\boldsymbol{D}}_{\sigma^2}.$$

Let B_{ij} be the $(i,j)^{th}$ element of $\boldsymbol{B}$ then

$$\begin{aligned}\frac{\partial B_{ij}}{\partial \sigma^2} = \frac{1}{2}&\exp\left(\bar{\alpha}_i\right)\frac{\partial \bar{\alpha}_i}{\partial \sigma^2}\exp\left(\bar{\alpha}_j\right)\bar{M}_{ij}^2 + \exp\left(\bar{\alpha}_i\right)\exp\left(\bar{\alpha}_j\right)\frac{\partial \bar{\alpha}_j}{\partial \sigma^2}\bar{M}_{ij}^2 \\ &- \exp\left(\bar{\alpha}_i\right)\exp\left(\bar{\alpha}_j\right)\bar{M}_{ij}\boldsymbol{l}_i^T\bar{\boldsymbol{H}}^{-2}\boldsymbol{l}_j - \exp\left(\bar{\alpha}_i\right)\exp\left(\bar{\alpha}_j\right)\bar{M}_{ij}\boldsymbol{l}_i^T\bar{\boldsymbol{H}}^{-1}\boldsymbol{L}\bar{\boldsymbol{D}}_{\sigma^2}\boldsymbol{L}^T\bar{\boldsymbol{H}}^{-1}\boldsymbol{l}_j\end{aligned}$$

and the derivative of the full matrix $\boldsymbol{\beta}$ is

$$\begin{aligned}\frac{\partial \boldsymbol{B}}{\partial \sigma^2} = \frac{1}{2}&\left(\left(\exp\left(\bar{\boldsymbol{\alpha}}\right)\circ\frac{\partial \bar{\boldsymbol{\alpha}}}{\partial \sigma^2}\right)\exp\left(\bar{\boldsymbol{\alpha}}\right)^T + \exp\left(\bar{\boldsymbol{\alpha}}\right)\left(\exp\left(\bar{\boldsymbol{\alpha}}\right)\circ\frac{\partial \bar{\boldsymbol{\alpha}}}{\partial \sigma^2}\right)^T\right)\circ\bar{\boldsymbol{M}}\circ\bar{\boldsymbol{M}} \\ &- \exp\left(\bar{\boldsymbol{\alpha}}\right)\exp\left(\bar{\boldsymbol{\alpha}}\right)^T\circ\bar{\boldsymbol{M}}\circ\left(\boldsymbol{L}^T\bar{\boldsymbol{H}}^{-2}\boldsymbol{L} + \bar{\boldsymbol{M}}\bar{\boldsymbol{D}}_{\sigma^2}\bar{\boldsymbol{M}}\right).\end{aligned}$$

The derivative of the conditional expected information is the sum of the two component derivatives. □

Theorem 10.6. *The score equations for σ^2 and $2\phi^2$ are*

$$\begin{aligned}\frac{\partial \ell}{\partial \sigma^2} &= -\frac{1}{2}\operatorname{tr}\left(\bar{\boldsymbol{\Lambda}}^{-1}\frac{\partial \bar{\boldsymbol{\Lambda}}}{\partial \sigma^2}\right) - \frac{1}{2}\operatorname{tr}\left(\bar{\boldsymbol{H}}^{-1}\right) - \frac{1}{2}\operatorname{tr}\left(\boldsymbol{L}^T\bar{\boldsymbol{H}}^{-1}\boldsymbol{L}\bar{\boldsymbol{D}}_{\sigma^2}\right) + \frac{1}{2}\boldsymbol{y}^T\bar{\boldsymbol{H}}^{-2}\boldsymbol{y} \\ &\quad + \frac{1}{2}\boldsymbol{y}^T\bar{\boldsymbol{H}}^{-1}\boldsymbol{L}\bar{\boldsymbol{D}}_{\sigma^2}\boldsymbol{L}^T\bar{\boldsymbol{H}}^{-1}\boldsymbol{y} - \frac{1}{2\phi^2}\sum_{i=1}^{q}\exp\left(\bar{\alpha}_i\right)\frac{\partial \bar{\alpha}_i}{\partial \sigma^2} + \sum_{i=1}^{q}\frac{\partial \bar{\alpha}_i}{\partial \sigma^2} \\ \frac{\partial \ell}{\partial 2\phi^2} &= -\frac{q}{2\phi^2} - \frac{1}{2}\operatorname{tr}\left(\bar{\boldsymbol{\Lambda}}^{-1}\frac{\partial \bar{\boldsymbol{\Lambda}}}{\partial 2\phi^2}\right) - \frac{1}{2}\operatorname{tr}\left(\boldsymbol{L}^T\bar{\boldsymbol{H}}^{-1}\boldsymbol{L}\bar{\boldsymbol{D}}_{2\phi^2}\right) + \frac{1}{2}\boldsymbol{y}^T\bar{\boldsymbol{H}}^{-1}\boldsymbol{L}\bar{\boldsymbol{D}}_{2\phi^2}\boldsymbol{L}^T\bar{\boldsymbol{H}}^{-1}\boldsymbol{y} \\ &\quad + \frac{1}{4\phi^4}\sum_{i=1}^{q}\exp\left(\bar{\alpha}_i\right) - \frac{1}{2\phi^2}\sum_{i=1}^{q}\exp\left(\bar{\alpha}_i\right)\frac{\partial \bar{\alpha}_i}{\partial 2\phi^2} + \sum_{i=1}^{q}\frac{\partial \bar{\alpha}_i}{\partial 2\phi^2}.\end{aligned}$$

Proof. Using the results from Theorems 10.3, 10.4 and 10.2, the derivatives of the log-likelihood in Theorem 10.1 are readily obtained. □

10.5 Estimation

The maximum likelihood estimates of σ^2 and $2\phi^2$ are obtained from maximising the approximate log-likelihood in Theorem 10.1 using the score equations in Theorem 10.6. The maximum of the approximate log-likelihood is obtained using a slightly altered form of the Newton-Raphson (NR) method.

The standard NR method requires the Hessian of the log-likelihood to be calculated at each updated position of the parameters. An analytic form for the Hessian has not been given as the derivation would be extremely difficult, requiring second derivatives of $\bar{\boldsymbol{\alpha}}$, $\bar{\boldsymbol{\Lambda}}_o$ and/or $\bar{\boldsymbol{\Lambda}}_e$. The analytic form, even if it were available, would require a substantial amount of computations on potentially large matrices and so it is likely to be computationally prohibitive to calculate. Numerical methods such as the Quasi-Newton method (e.g. Stoer & Bulirsch, 1993, page 317) could be used. This method is employed by S-PLUS' function `nlminb` (A. T. & T. Bell Laboratories, 1984). However, the simpler method of approximating the Hessian at each updated estimate is employed here.

The approximation is found by taking the matrix of first forward differences of the score equations (a similar idea to the numerical methods for estimating the penalised LASSO objective in Chapter 6). Let

$$\boldsymbol{A} = \begin{pmatrix} A_{\sigma^2\sigma^2} & A_{\sigma^2 2\phi^2} \\ A_{2\phi^2\sigma^2} & A_{2\phi^2 2\phi^2} \end{pmatrix}$$

be the approximate Hessian of the log-likelihood for parameters σ^2 and $2\phi^2$. The individual elements of $\boldsymbol{A}$ at position $(\sigma^2, 2\phi^2)$ are given by

$$\begin{aligned}
A_{\sigma^2\sigma^2} &= \frac{\frac{\partial \ell}{\partial \sigma^2}(\sigma^2+\varepsilon, 2\phi^2) - \frac{\partial \ell}{\partial \sigma^2}(\sigma^2, 2\phi^2)}{\varepsilon} \\
A_{2\phi^2 2\phi^2} &= \frac{\frac{\partial \ell}{\partial 2\phi^2}(\sigma^2, 2\phi^2+\varepsilon) - \frac{\partial \ell}{\partial 2\phi^2}(\sigma^2, 2\phi^2)}{\varepsilon} \\
A^*_{\sigma^2 2\phi^2} &= \frac{\frac{\partial \ell}{\partial \sigma^2}(\sigma^2, 2\phi^2+\varepsilon) - \frac{\partial \ell}{\partial \sigma^2}(\sigma^2, 2\phi^2)}{\varepsilon} \\
A^*_{2\phi^2\sigma^2} &= \frac{\frac{\partial \ell}{\partial 2\phi^2}(\sigma^2+\varepsilon, 2\phi^2) - \frac{\partial \ell}{\partial 2\phi^2}(\sigma^2, 2\phi^2)}{\varepsilon} \\
A_{\sigma^2 2\phi^2} &= A_{2\phi^2\sigma^2} = \frac{1}{2}(A^*_{\sigma^2 2\phi^2} + A^*_{2\phi^2\sigma^2})
\end{aligned}$$

where ε is some small positive number. The quantities $A^*_{2\phi^2\sigma^2}$ and $A^*_{\sigma^2 2\phi^2}$ should be equal if the approximation is accurate, as the second (exact) derivative is invariant to the order of differentiation. However, they cannot be assumed equal in the approximation and the average is used as it is likely to be more stable.

This approximation requires multiple calculations of the score equations and hence multiple evaluations of $\bar{\boldsymbol{\alpha}}$. Prior to evaluation of the approximate Hessian, $\bar{\boldsymbol{\alpha}}$ is calculated at $(\sigma^2, 2\phi^2)$ and these provide excellent starting values for the calculation of $\bar{\boldsymbol{\alpha}}$ at $(\sigma^2+\varepsilon, 2\phi^2)$ and $(\sigma^2, 2\phi^2+\varepsilon)$. The calculation of the approximate Hessian requires little additional computation above calculation of the score equations.

The choice of ε depends heavily on the convergence tolerance when calculating $\bar{\boldsymbol{\alpha}}$ using the scoring method in Section 10.3.1. In particular, if ε is relatively small compared to the tolerance level, then the approximate Hessian for the log-likelihood can have positive diagonal elements. This would appear to be incorrect as the log-likelihood surface is likely to be concave (e.g. Figure 10.1). In such situations, the NR routine may diverge. To overcome this, a small tolerance value should be used. In the simulations in this chapter, the tolerance was set to 10^{-6} and ε was chosen to be 0.01. These have been found to work well for the analyses in this chapter but they may have to be changed for any particular data set.

10.6 Performance of Estimation Methods

The estimation methods presented in this chapter are evaluated via simulation. There are two aspects to the estimation that can be examined. The first is to see if the approximate log-likelihoods (using the observed and expected information for $\boldsymbol{\alpha}$) are unbiased for the dispersion parameters. That is, if the estimation methods are efficacious. The second is to investigate how the new methods compare to the estimation method presented in Chapter 7.

10.6.1 *Efficacy*

Efficacy was examined using 9 closely related simulations. Specifically, for each combination of $\sigma^2 = (2.0, 1.0, 0.5)$ and $2\phi^2 = (2.0, 1.0, 0.5)$, data sets were simulated and the dispersion parameters were estimated using the approximate log-likelihoods with both the observed and conditional expected information matrices for $\boldsymbol{\alpha}$. A total of 2000 data sets of each dispersion parameter combination were simulated and analysed using both the conditional expected information for $\boldsymbol{\alpha}$, and the observed information for $\boldsymbol{\alpha}$. For each simulation the number of observations was $n = 50$ and the number of LASSO effects was $q = 20$. The dispersion parameter estimates were recorded for each data set. A summary of the results are presented in Table 10.1.

Table 10.1: Means (± standard errors) of the estimates from the alternative LASSO random effects model approximate likelihoods. The common simulation values were: $n = 50$ and $q = 20$. 2000 simulations were used to generate each of these statistics.

Simulation Values		Expected Information		Observed Information	
σ^2	$2\phi^2$	$\hat{\sigma}^2$	$2\hat{\phi}^2$	$\hat{\sigma}^2$	$2\hat{\phi}^2$
2.0	2.0	1.939 (0.016)	1.558 (0.022)	1.886 (0.015)	2.148 (0.029)
	1.0	1.925 (0.015)	0.790 (0.011)	1.850 (0.015)	1.109 (0.016)
	0.5	1.968 (0.017)	0.398 (0.006)	1.873 (0.016)	0.566 (0.009)
1.0	2.0	0.973 (0.008)	1.552 (0.022)	0.956 (0.008)	2.107 (0.030)
	1.0	0.980 (0.008)	0.792 (0.012)	0.953 (0.008)	1.091 (0.016)
	0.5	0.981 (0.008)	0.391 (0.006)	0.942 (0.008)	0.550 (0.008)
0.5	2.0	0.481 (0.004)	1.513 (0.022)	0.476 (0.004)	2.031 (0.029)
	1.0	0.480 (0.004)	0.761 (0.011)	0.471 (0.004)	1.032 (0.014)
	0.5	0.485 (0.004)	0.389 (0.006)	0.472 (0.004)	0.537 (0.008)

The approximate log-likelihood based on the conditional expected information for $\boldsymbol{\alpha}$ produced estimates for $2\phi^2$ that were substantially biased downward. However, the estimates from the approximate log-likelihood based on the observed information for $\boldsymbol{\alpha}$ were far less biased. However, they were slightly biased downwards for σ^2 and upwards for $2\phi^2$. Clearly, the observed information approximation should be preferred.

The bias in both parameter estimates from the observed information for $\boldsymbol{\alpha}$ was alleviated when σ^2 was larger than $2\phi^2$. This implies that in a practical situations it may be beneficial to re-scale the explanatory variables (using a linear transformation) after an initial analysis has been performed.

The approximate log-likelihood based on the observed information matrix for $\boldsymbol{\alpha}$ appeared to be superior to that based on the conditional expected information for $\boldsymbol{\alpha}$. For the remainder of the thesis the latter approximation is disregarded.

10.6.2 *Comparison with Bootstrap Adjustment Method*

The dispersion parameter estimates found using the approximate log-likelihood was compared with those from the bootstrap adjusted score estimation method. A total of 100 data sets were simulated with $n = 100$, $q = 35$, $\sigma^2 = 1$ and $2\phi^2 = 1$. Each data set was analysed using both methods, and the dispersion parameters along with the time taken to finish estimation were recorded. A convergence tolerance for the dispersion parameters of 0.001 was used for both methods. The results for the dispersion parameter estimates are presented in Figure 10.2. If there was complete agreement in the estimates from the two methods then observed points should lie on the $y = x$ line.

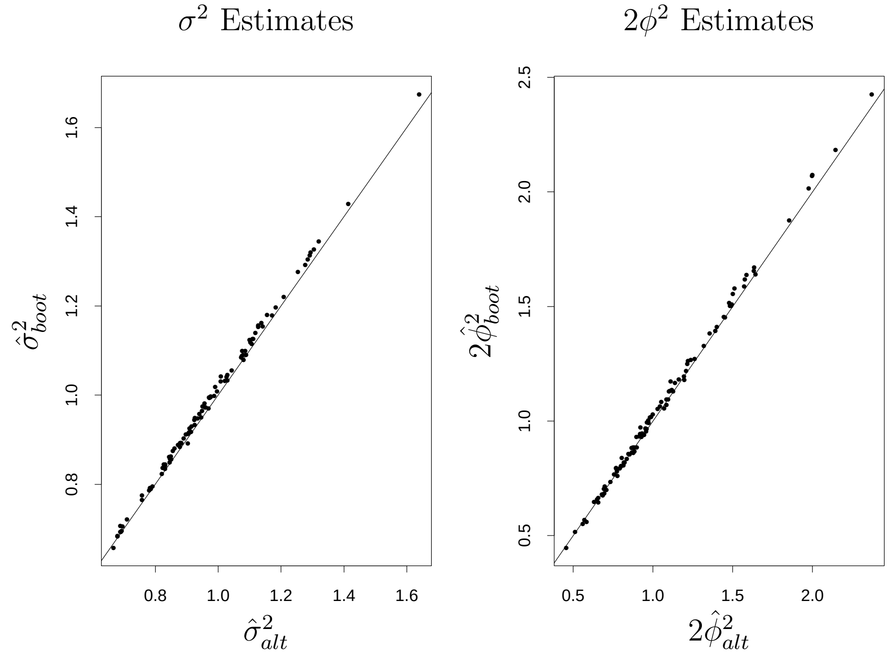

Figure 10.2: Plot of the estimates of σ^2 and $2\phi^2$ obtained from the bootstrap adjusted score method versus those obtained from the approximate log-likelihood in this Chapter. The solid line is $y = x$.

There was good agreement between the two methods (Figure 10.2). It appears that estimation using bootstrap adjusted scores may lead to slightly larger estimates for both σ^2

and $2\phi^2$. The mean (± standard error) for the estimates based on the methods in Chapter 7 were $\hat{\sigma}^2 = 0.991$ (± 0.020) and $2\hat{\phi}^2 = 1.073$ (± 0.038) and corresponding values for the methods developed in this chapter were $\hat{\sigma}^2 = 0.977$ (± 0.019) and $2\hat{\phi}^2 = 1.060$ (± 0.036). Some of the discrepancies are likely to have occurred because of the stochastic element in the estimation for the bootstrap adjusted scores. However, this should not explain the small differences in the means.

Estimation was performed on a Linux computer with processor speed 2.2 GHz and 2.0 Gb of RAM. The mean CPU time for estimation using the methods in Chapter 7 was 46.11 minutes (±1.82 minutes) and the mean CPU time for estimation using the method described in this chapter was 1.60 minutes (±0.03 minutes). The difference is enormous and given that the estimates do not differ substantially, it shows that the log-likelihood based on the alternative model should be used in practice. However, the difference in CPU time will change depending on the attributes of the data, in particular the sample size n and the number of LASSO effects q. Increasing n should not increase the estimation time for the bootstrap adjusted score method substantially. It will increase estimation time for the alternative random effects model approximate log-likelihood as $\boldsymbol{H}$ has to be formed and inverted. Increasing q will increase estimation time for both methods.

10.7 Prediction of LASSO Random Effects

Up to this point, prediction for the LASSO random effects has been performed using the suggestion from Tibshirani (1996). That is, the predictors were chosen to be the mode of the predictive distribution $f(\boldsymbol{\beta}|\boldsymbol{y})$. If the criterion for prediction is to minimise the mean squared error of prediction (MSE), then the expected value of the predictive distribution should be used (Searle et al., 1992, Chapter 7). This predictor is the best predictor (BP), which is likely to be non-linear in non-normal models. If the predictor is constrained to be a linear function of the data $\boldsymbol{y}$ then a third type of predictor is available, the best linear predictor (BLP).

The standard LASSO, BP and BLP predictors are different summaries of the predictive distribution marginal to $\boldsymbol{\delta}$. However, the alternative LASSO model also suggests that conditioning on $\boldsymbol{\delta}$ may also lead to useful predictions as the conditional predictive distribution is normal and standard normal theory applies. However, the choice of which value of $\boldsymbol{\delta}$ to choose is not obvious. Sensible choices are limited due to practical considerations; two are considered here. The first conditions on $\boldsymbol{\delta} = \bar{\boldsymbol{\delta}}$ where $\bar{\boldsymbol{\delta}}$ is the maximum of the predictive distribution $f(\boldsymbol{\delta}|\boldsymbol{y})$. The second conditions on $\boldsymbol{\delta} = \bar{\boldsymbol{\delta}}_{\boldsymbol{\alpha}} = \exp(\bar{\boldsymbol{\alpha}})$ where $\bar{\boldsymbol{\alpha}}$ is the maximum of the predictive distribution of $f(\boldsymbol{\alpha}|\boldsymbol{y})$ found during the last iteration of dispersion parameter estimation. The values $\bar{\boldsymbol{\delta}}$ and $\exp(\bar{\boldsymbol{\alpha}})$ will not be equal due to the constrained solution space of $\boldsymbol{\delta}$. It is expected that some of the $\bar{\delta}_i$ may be identically zero.

All the predictors considered require knowledge of the dispersion parameters. These are unknown and are replaced using their estimates. The resulting predictors should formally called empirical predictions but in this section the 'empirical' adjective is dropped.

Descriptions of how to calculate all predictors are now given with the exception of the predictive mode (described in Chapter 6).

10.7.1 *Best Linear Predictor*

Calculation of the BLP requires only knowledge of the mean and variance of the joint distribution of the random LASSO effects $\boldsymbol{\beta}$ and the observed outcomes $\boldsymbol{y}$. It is convenient to use the random effects model in (7.1.1) where the random LASSO effects possess a centred double exponential distribution. The mean for both the LASSO effects and the outcomes is the zero vector. The variance and covariances are $\text{var}\left(\boldsymbol{y}\right) = \sigma^2 \boldsymbol{H}_a$, $\text{var}\left(\boldsymbol{\beta}\right) = 2\phi^2 \boldsymbol{I}$ and $\text{cov}\left(\boldsymbol{y}, \boldsymbol{\beta}\right) = 2\phi^2 \boldsymbol{L}$ where $\boldsymbol{H}_a = \sigma^2 \boldsymbol{I} + 2\phi^2 \boldsymbol{L}\boldsymbol{L}^T$ and $\boldsymbol{I}$ is the identity matrix of appropriate dimension. The BLP and its associated variance are (Searle et al., 1992)

$$\begin{aligned} \mathrm{E}\left(\boldsymbol{\beta}|\boldsymbol{y}\right) &= 2\phi^2 \boldsymbol{L}^T \boldsymbol{H}_a^{-1} \boldsymbol{y} \\ \text{var}\left(\boldsymbol{\beta}|\boldsymbol{y}\right) &= 2\phi^2 \boldsymbol{I} - 4\phi^4 \boldsymbol{L}^T \boldsymbol{H}_a^{-1} \boldsymbol{L} \end{aligned} \tag{10.7.1}$$

If the predictive distribution is approximately normal then the BLP predictor should approximate the BP well.

10.7.2 *Best Predictor*

There is not a closed form expression for the LASSO BP. Here, it is calculated numerically using importance sampling (e.g. Gelman et al., 1995). In particular,

$$\begin{aligned} \mathrm{E}\left(\boldsymbol{\beta}|\boldsymbol{y}\right) &= \int_{\mathbb{R}^q} \boldsymbol{\beta} f(\boldsymbol{\beta}|\boldsymbol{y}) \partial \boldsymbol{\beta} \\ &= \int_{\mathbb{R}^q} \boldsymbol{\beta} \frac{f(\boldsymbol{\beta}|\boldsymbol{y})}{g(\boldsymbol{\beta})} g(\boldsymbol{\beta}) \partial \boldsymbol{\beta} \\ &= \mathrm{E}\left(\boldsymbol{\beta} \frac{f(\boldsymbol{\beta}|\boldsymbol{y})}{g(\boldsymbol{\beta})}\right) \\ &\approx \frac{1}{B} \sum_{i=1}^{B} \boldsymbol{\beta}_i \frac{f(\boldsymbol{\beta}_i|\boldsymbol{y})}{g(\boldsymbol{\beta}_i)} \end{aligned}$$

where $g(\boldsymbol{\beta})$ is a proposal distribution that approximates the predictive distribution reasonably well and $\boldsymbol{\beta}_i$ is the i^{th} of B samples from the distribution $g(\boldsymbol{\beta})$. The ratios $w_i = f(\boldsymbol{\beta}_i|\boldsymbol{y})/g(\boldsymbol{\beta}_i)$ are called 'importance weights'.

The BP approximation requires evaluation of the predictive distribution. An application of Bayes' law gives this distribution to be $f(\boldsymbol{\beta}|\boldsymbol{y}) = f(\boldsymbol{y}|\boldsymbol{\beta})f(\boldsymbol{\beta})/f(\boldsymbol{y})$. There is no exact method to calculate the marginal distribution but analytic approximations using the methods described in Section 10.3 could be used. However, this could be computationally prohibitive due to the $n \times n$ matrix $\boldsymbol{H}$ being formed and inverted for every sample. The marginal distribution can also be approximated using the same set of samples generated in importance sampling. Using a similar derivation to the approximation of $\mathrm{E}\left(\boldsymbol{\beta}|\boldsymbol{y}\right)$ the marginal approximation is

$$f(\boldsymbol{y}) = \int_{\mathbb{R}^q} f(\boldsymbol{y}, \boldsymbol{\beta}) \partial \boldsymbol{\beta} = \mathrm{E}\left(\frac{f(\boldsymbol{y}, \boldsymbol{\beta})}{g(\boldsymbol{\beta})}\right) \approx \frac{1}{B} \sum_{i=1}^{B} w_i^*.$$

where $w_i^* = f(\boldsymbol{\beta}_i, \boldsymbol{y})/g(\boldsymbol{\beta}_i)$. The BP is then

$$\begin{aligned} \mathrm{E}\left(\boldsymbol{\beta}|\boldsymbol{y}\right) &\approx \frac{\frac{1}{B}\sum_{i=1}^{B} \boldsymbol{\beta}_i w_i^*}{\frac{1}{B}\sum_{j=1}^{B} w_j^*} \\ &= \frac{\sum_{i=1}^{B} \boldsymbol{\beta}_i w_i^*}{\sum_{j=1}^{B} w_j^*}. \end{aligned} \tag{10.7.2}$$

A multivariate normal distribution with mean and variance defined by BLP (10.7.1) is used as the proposal distribution. This may not be the most appropriate proposal distribution but it is easily computed.

Gelman et al. (1995) suggest that importance sampling can give imprecise results if the proposal distribution is not appropriately matched to the predictive distribution. Thus, good results can be obtained if the large importance weights occur with high probability. A good proposal distribution generates importance weights that have a left skewed distribution. A plot of the frequencies of the (log) importance weights provides a check (Gelman et al., 1995) of the adequacy of the proposal distribution. As an example for the prediction problem for the LASSO random effects model is given in Figure 10.3. It is neither obviously left nor right skewed and it is assumed that good results can be obtained (for a sufficiently large B) using this proposal distribution.

Other proposal distributions, such as multivariate t-distributions with low degrees of freedom, have also been investigated as these place more mass in the tails of the distribution. Proposal distributions of this type did not improve the distribution of the importance ratio significantly (results not shown). Hence, they were not used. Also, the proposal distribution with mean equal to the LASSO predictions (the predictive distribution's mode) rather than the BLP was considered. Once again, this did not improve the distribution of the importance weights.

10.7.3 *Conditional Prediction*

An appealing feature of the random effects model in (10.2.1) is that conditional on $\boldsymbol{\delta}$, the predictive distribution is normal. This observation suggests that conditioning on $\boldsymbol{\delta}$ is a convenient method for prediction. In this chapter, these predictors are called *conditional* best predictors (CBP). Prediction of this type can be partially justified by observing that it is the prediction of a random effect that is required (not the effect's variance). However, this argument does not provide full justification as the distribution of the random effects changes once conditioning occurs. Another shortcoming of conditional prediction is choosing the

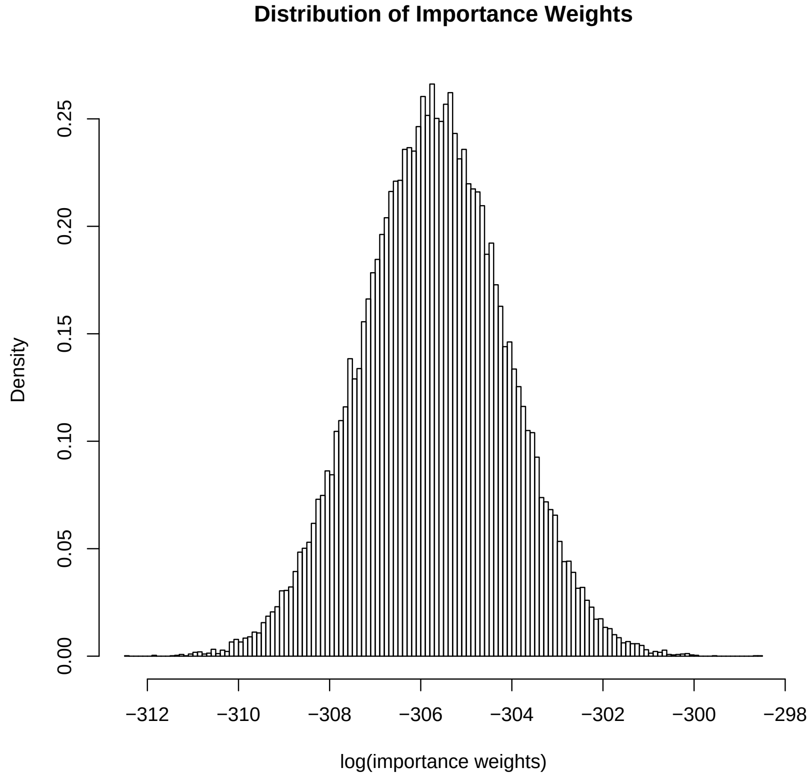

Figure 10.3: Distribution of the log of the observed importance weights.

value of $\boldsymbol{\delta}$ to condition on. Two choices are presented and investigated here. The first is the maximum of the predictive distribution $\bar{\boldsymbol{\delta}} = \max f(\boldsymbol{\delta}|\boldsymbol{y}) = \max f(\boldsymbol{\delta}, \boldsymbol{y})$. The second is chosen because of convenience rather than any particular statistical reason. It is $\bar{\boldsymbol{\delta}}_{\boldsymbol{\alpha}} = \exp(\bar{\boldsymbol{\alpha}})$ where $\bar{\boldsymbol{\alpha}}$ is the maximum of the predictive distribution $f(\boldsymbol{\alpha}|\boldsymbol{y}) \propto f(\boldsymbol{\alpha}, \boldsymbol{y})$. The values $\bar{\boldsymbol{\alpha}}$ have already been calculated in the last iteration of the dispersion parameter estimation.

The CBP estimates are

$$\begin{aligned} \text{CBP}_{\bar{\boldsymbol{\delta}}} &= \bar{\boldsymbol{D}}_{\boldsymbol{\delta}} \boldsymbol{L}^T \bar{\boldsymbol{H}}_{\boldsymbol{\delta}}^{-1} \boldsymbol{y} \qquad \text{and} \\ \text{CBP}_{\bar{\boldsymbol{\alpha}}} &= \bar{\boldsymbol{D}}_{\boldsymbol{\alpha}} \boldsymbol{L}^T \bar{\boldsymbol{H}}_{\boldsymbol{\alpha}}^{-1} \boldsymbol{y} \end{aligned}$$

where $\bar{\boldsymbol{D}}_{\boldsymbol{\delta}} = \text{diagv}\left(\bar{\boldsymbol{\delta}}\right)$, $\bar{\boldsymbol{H}}_{\boldsymbol{\delta}} = \sigma^2 \boldsymbol{I} + \boldsymbol{L}\bar{\boldsymbol{D}}_{\boldsymbol{\delta}}\boldsymbol{L}^T$, $\bar{\boldsymbol{D}}_{\boldsymbol{\alpha}} = \text{diagv}(\exp(\bar{\boldsymbol{\alpha}}))$ and $\bar{\boldsymbol{H}}_{\boldsymbol{\alpha}} = \sigma^2 \boldsymbol{I} + \boldsymbol{L}\bar{\boldsymbol{D}}_{\boldsymbol{\alpha}}\boldsymbol{L}^T$.

The conditional predictors are fundamentally different from the marginal predictors. It is expected that their behaviour may also be different. In particular, there is no guarantee that the total variance of $(\boldsymbol{\beta}|\boldsymbol{\delta} = \bar{\boldsymbol{\delta}})$ or $(\boldsymbol{\beta}|\boldsymbol{\delta} = \exp(\bar{\boldsymbol{\alpha}}))$ is similar to that of $\boldsymbol{\beta}$. Hence, the predictors may be subjected to heavier (or lighter) constraints. One way around this might be to use an unbiased predictor for $\boldsymbol{\delta}$. However, it is not clear how this could be done.

Estimation of $\boldsymbol{\delta}$

Estimation of $\bar{\boldsymbol{\alpha}}$ has already been described (Section 10.3.1). A description of the method of estimation for $\bar{\boldsymbol{\delta}}$ is now given. This method is not ideal as it can be computationally expensive and can fail to converge if q is large. However, it easily suffices for the purposes of this thesis where it is only used for comparisons of predictors. More work will be required if the conditional predictors are considered to be desirable.

Estimation requires the maximisation of the joint distribution, which is equivalent to maximisation of its exponent

$$Q(\boldsymbol{\delta}) = -\frac{1}{2}\log|\boldsymbol{H}_{\boldsymbol{\delta}}| - \frac{1}{2}\boldsymbol{y}^T\boldsymbol{H}_{\boldsymbol{\delta}}^{-1}\boldsymbol{y} - \frac{1}{2\phi^2}\sum_{i=1}^{q}\delta_i.$$

Since δ_i $(1 \leq i \leq q)$ is an exponential random variable, the maximum $\bar{\delta}_i$ must be greater than, or equal to zero. This implies that the solution space is the positive orthant and the solution to the constrained problem may lie on the boundary (in practice it often does). There are a number of numerical methods that could be employed, for example the Newton-Raphson (NR) and the quasi-Newton algorithms. These methods seem to be suitable for small problems ($q \lessapprox 10$, $n = 100$) but for larger problems both these methods are unstable and tend to diverge (or not converge). A slightly modified version of the Lindley-Smith algorithm (Lindley & Smith, 1972) has been found to behave much better. It is still not suitable for larger problems ($q \gtrapprox 50$, $n = 100$) but it does provide a method for comparison of the different predictors of the LASSO random effects.

The estimation procedure requires the first and second derivatives of $Q(\boldsymbol{\delta})$ with respect to $\boldsymbol{\delta}$. The derivation of these derivatives follows closely to that of the first and second derivatives of $Q(\boldsymbol{\alpha})$ with respect to $\boldsymbol{\alpha}$ in Section 10.3.1. For this reason, a full derivation for the derivatives is not given. The derivatives are

$$\begin{aligned}\frac{\partial Q(\boldsymbol{\delta})}{\partial\boldsymbol{\delta}} &= -\frac{1}{2}\text{vecd}\,(\boldsymbol{M}) + \frac{1}{2}\boldsymbol{w}\circ\boldsymbol{w} - \frac{1}{2\phi^2}\boldsymbol{1}\\ &= U(\boldsymbol{\delta}) \qquad\qquad \text{and}\\ \frac{\partial^2 Q(\boldsymbol{\delta})}{\partial\boldsymbol{\delta}\partial\boldsymbol{\delta}^T} &= \frac{1}{2}\boldsymbol{M}\circ\boldsymbol{M} - \boldsymbol{w}\boldsymbol{w}^T\circ\boldsymbol{M}\\ &= \boldsymbol{\Lambda}_{\boldsymbol{\delta}}.\end{aligned}$$

The Lindley-Smith algorithm updates all the individual δ_i in turn by solving the score equation for the single effect conditional on all other effects equal to their latest estimate. Estimation stops when all the score equations have been solved. There is no closed form for updating the individual δ_i and so the NR algorithm is used. That is, the $(j+1)^{th}$ iteration's estimate of δ_i is

$$\delta_i^{(j+1)} = \delta_i^{(j)} + s\boldsymbol{\Lambda}_{\boldsymbol{\delta},ii}^{-1}U_i(\boldsymbol{\delta})$$

where $\boldsymbol{\Lambda}_{\boldsymbol{\delta},ii}$ is the i^{th} diagonal of $\boldsymbol{\Lambda}_{\boldsymbol{\delta}}$, $U_i(\boldsymbol{\delta})$ is the i^{th} element of $U(\boldsymbol{\delta})$, $0 < s \leq 1$ and the score and information are evaluated at $\boldsymbol{\delta} = \bar{\boldsymbol{\delta}}^{(j)}$. Updating each δ_i is iterative and the algorithm stops when the i^{th} score equation is solved to within a predefined tolerance. A small step size was chosen as default ($s = 0.1$) to guard against gross over-shooting of the maximum, which can occur reasonably frequently.

In this constrained situation, the estimation algorithm was started with $\bar{\boldsymbol{\delta}} = \mathbf{0}$ as steps away from zero tend to be more stable (less dramatic) than steps towards it. All the effects whose current estimate is zero and whose score equation less than zero can be ignored as any update would be negative. Of the remaining effects, the effect with largest absolute score (the most violated effect) is chosen for updating. This choice of variable differs from the standard Lindley-Smith algorithm, which cycles through effects in a predefined order. The updating routine is repeated until all the non-zero effects have zero scores and all others have negative scores.

It is advisable to choose an equivalent (or stricter) tolerance level for updating each individual effect than for overall convergence. If this were not the case, even the effect most recently updated may not pass the overall convergence criterion. The defaults chosen here are 0.001 for both overall and individual effect convergence.

10.7.4 *Example of Different Predictors*

To illustrate the attributes of the different types of predictors a single data set was simulated and analysed. The data set contained $n = 100$ observations, $q = 50$ explanatory variables, residual variance of $\sigma^2 = 10$ and random LASSO effect variance $2\phi^2 = 1$. These values were chosen to induce some LASSO predictions to be identically zero. Using the estimation methods presented in Section 10.5, the dispersion parameters were estimated to be $\hat{\sigma}^2 = 8.11$ and $2\hat{\phi}^2 = 1.08$.

The best predictor was calculated using 50 000 importance samples. This is excessive (a small fraction of this should suffice) but there was concern about the curse of dimensionality before performing this prediction.

The different predictors and the realised random effects are plotted against each other in a scatter plot matrix in Figure 10.4. For this single data set, there is good general agreement between the different types of predictors. Two prediction methods (LASSO and CBP$_{\bar{\boldsymbol{\delta}}}$ exhibited the ability to estimate identically zero effects. CBP$_{\bar{\boldsymbol{\delta}}}$ estimated more zero effects ($q_{\mathcal{Z}} = 23$) than the LASSO method ($q_{\mathcal{Z}} = 9$). By definition, the mean variance of the marginal effects was $2\hat{\phi}^2 = 1.08$. The mean variance of the conditional distributions were 0.95 for CBP$_{\tilde{\boldsymbol{\alpha}}}$ and 0.30 for CBP$_{\bar{\boldsymbol{\delta}}}$. Both of these are considerably lower than the marginal counterparts.

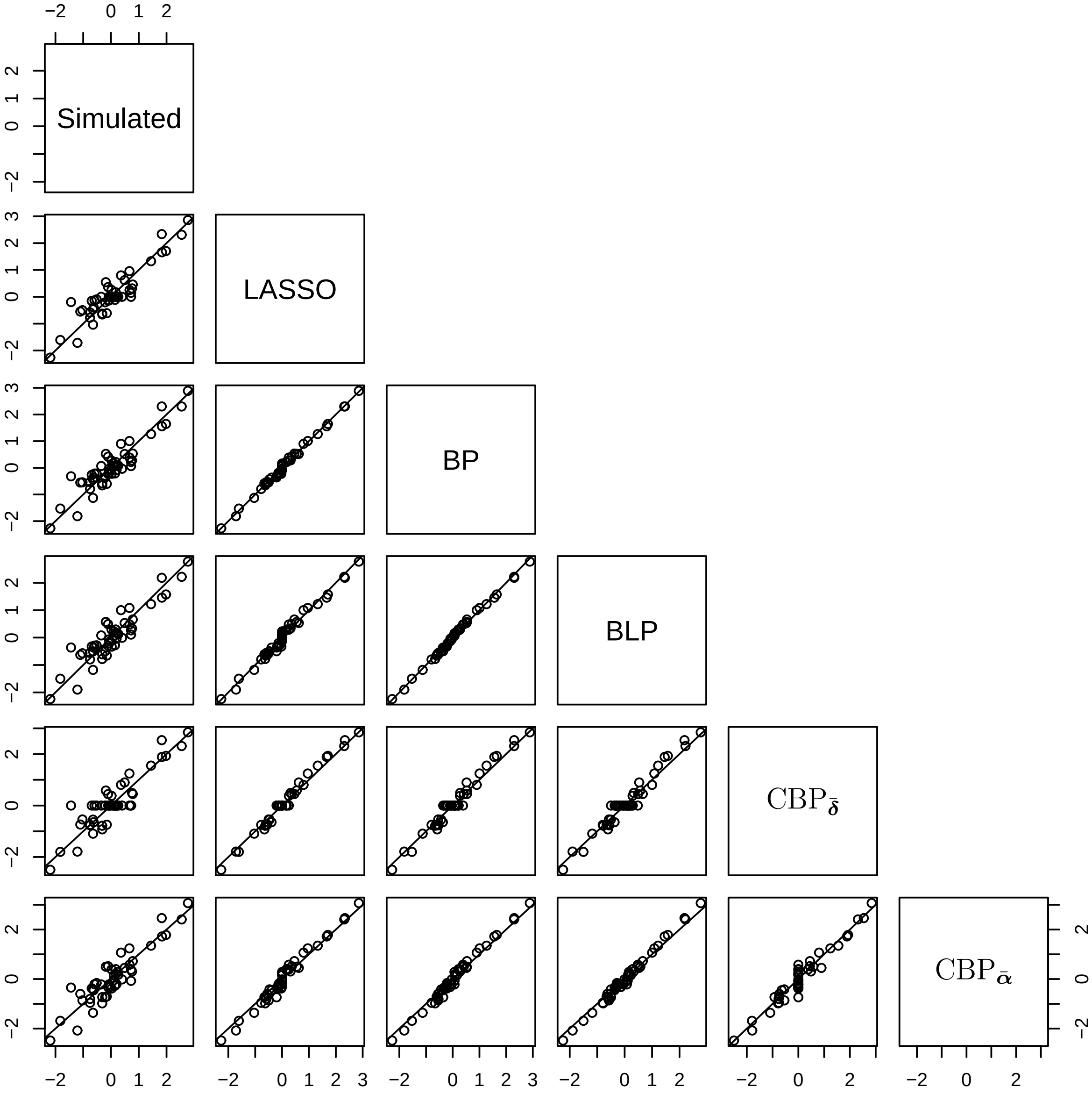

Figure 10.4: Scatter plot matrix of different predictors estimated from the same data. Solid line is $y = x$.

10.7.5 *Performance of Predictors*

By definition, the best predictor should have lowest MSE of all the (marginal) predictors. However, there is a stochastic element to calculating the BP and hence it may not exhibit this property in practice. Also, it is interesting to know how the standard LASSO predictor, the best linear predictor and both conditional predictors perform in relation to the BP and to each other. A simulation study was performed to investigate this.

The simulation study consisted of simulating and analysing 500 data sets with attributes: $n = 100$, $q = 15$, $\sigma^2 = 10$ and $2\phi^2 = 1$. For each simulated data set, the MSE and the number of identically zero predictions were recorded. The simulation results are given in Table 10.2. The average dispersion parameter estimates were 9.88 (±0.07) for σ^2 and 1.21 (±0.03) for $2\phi^2$. These follow the same pattern of bias found in the efficacy simulation study (Section 10.6).

Table 10.2: Average mean squared error (MSE), number of zero predictions ($q_{\mathcal{Z}}$) and average variance of predictors (MVP) for each of the different prediction methods. Parenthetic values are standard errors. Dashes in $q_{\mathcal{Z}}$ column signify no zero estimates.

Predictor	MSE	$q_{\mathcal{Z}}$	MVP
BP	0.1003 (0.0017)	-	1.21 (0.03)
LASSO	0.1038 (0.0018)	2.49 (0.08)	1.21 (0.03)
BLP	0.1041 (0.0017)	-	1.21 (0.03)
$\text{CBP}_{\bar{\boldsymbol{\alpha}}}$	0.1026 (0.0017)	-	1.04 (0.02)
$\text{CBP}_{\bar{\boldsymbol{\delta}}}$	0.1193 (0.0022)	5.35 (0.09)	0.53 (0.01)

The simulation shows that BP is the most statistically efficient predictor of the all predictors considered. The LASSO and BLP have similar prediction ability. Of the three marginal predictors, only the LASSO has the ability to generate identically zero predictions. The conditional predictors behave quite differently. The $\text{CBP}_{\bar{\boldsymbol{\delta}}}$ method generates more zero predictions than the LASSO method. The $\text{CBP}_{\bar{\boldsymbol{\alpha}}}$ method forms more accurate predictions than either the LASSO method or BLP. The average variance of the predictions for both conditional predictors were lower on average than the average $2\hat{\phi}^2$. In fact, for all 500 simulations the estimate of $2\hat{\phi}^2$ was higher than the average variance of the conditional predictors. This implies that there is, on average across effects, a larger constraint for the conditional predictors. In spite of this, the $\text{CBP}_{\bar{\boldsymbol{\alpha}}}$ predictor is still competitive with the marginal predictors.

In any practical situation it appears that using any of these predictors (except $\text{CBP}_{\bar{\boldsymbol{\delta}}}$) will provide useful predictions. However of these predictors, only the LASSO has the attractive ability to predict effects to be identically zero. It is expected to be a useful choice when predicting for models where interpretation is important.

10.8 Inference for Random LASSO Effects

In Chapter 7 inference was carried out for the model by performing the hypothesis tests with null hypothesis $\beta_i = 0$. This method was subsequently shown to be useful in determining the important effects from the rest. This test treats the effects as fixed even though they are not. The hypothesis test is not strictly appropriate for random effects as the effects' distribution is estimated from the data and not the individual effects themselves. The latter are predicted from the estimated predictive distribution.

Inference for random effects should be made from predictive probability statements. Of particular interest for the LASSO random model are the probabilities $\mathrm{P}(\beta_i > 0|\boldsymbol{y})$ and $\mathrm{P}(\beta_i < 0|\boldsymbol{y})$. If either of these probabilities are large then the explanatory variable is considered *statistically important*. Of course, the statistical significance of an effect does not necessarily imply biological/genetical significance. For biological/genetical significance the probabilities $\mathrm{P}(\beta_i = 0|\boldsymbol{y})$ and $\mathrm{P}(\beta_i \neq 0|\boldsymbol{y})$ would be extremely useful. However, under the random effects model the point probability is identically zero and hence these probabilities provide no insight using the LASSO random effects model.

Two methods to obtain these probabilities are now presented and compared. The first method uses importance sampling from the predictive distribution. The second method approximates the predictive distribution with a multivariate normal and probabilities are generated for each of the random effects.

The importance sampling approach requires the generation of importance weights and the probabilities are ratios of the sum of the importance weights in the relevant orthant to the total sum or the importance weights. This is now made clearer for the derivation of the approximation for $\mathrm{P}(\beta_i > 0|\boldsymbol{y})$. Define $\mathbb{R}_i^q$ be the positive half-space of $\mathbb{R}^q$, defined by $\beta_i > 0$, and the function $\mathrm{I}(\beta_i > 0|\boldsymbol{y})$ be the indicator function taking a value of 1 if $\beta_i > 0$ and 0 otherwise. Then

$$\begin{aligned}
\mathrm{P}(\beta_i > 0|\boldsymbol{y}) &= \int_{\mathbb{R}_i^q} f(\boldsymbol{\beta}|\boldsymbol{y})\partial\boldsymbol{\beta} \\
&= \int_{\mathbb{R}^q} \frac{\mathrm{I}(\beta_i > 0)f(\boldsymbol{\beta}|\boldsymbol{y})}{g(\boldsymbol{\beta})} g(\boldsymbol{\beta})\partial\boldsymbol{\beta} \\
&= \mathrm{E}\left(\frac{\mathrm{I}(\beta_i > 0)f(\boldsymbol{\beta}|\boldsymbol{y})}{g(\boldsymbol{\beta})}\right) \\
&= \mathrm{E}\left(\mathrm{I}(\beta_i > 0)w_j\right) \\
&\approx \frac{1}{B}\sum_{j=1}^{B}\Big(\mathrm{I}(\beta_i > 0)w_j\Big).
\end{aligned}$$

Paralleling the calculation of the best predictors, the predictive distribution's normalising factor is calculated using importance sampling. The approximate predictive probability is

$$\mathrm{P}(\beta_i > 0|\boldsymbol{y}) \approx \frac{\sum_{j=1}^{B}\mathrm{I}(\beta_i > 0)w_j^*}{\sum_{k=1}^{B} w_k^*}$$

where $w_j^* = f(\boldsymbol{\beta}_j, \boldsymbol{y})/g(\boldsymbol{\beta}_j)$ and $\boldsymbol{\beta}_j$ is the j^{th} sample from $g(\boldsymbol{\beta})$.

The second approach (based on approximating the predictive distribution by a multivariate normal) is simplistic, perhaps overly so. It requires little computation as the only extra calculations required are elementary matrix operations. The multivariate normal approximation is simply that defined by the mean and variance of the predictive distribution obtained using best linear prediction (10.7.1). The probability statements, $\mathrm{P}(\beta_i > 0|\boldsymbol{y})$, are made on the marginal univariate normal distributions.

Results from the single explanatory variable model in Chapter 5 show that when the predictor is large, the predictive distribution appears to be symmetrical and reasonably normal (see Theorem 5.2 and Figure 5.1). If this results also holds for the multiple explanatory variable situation, then the probability statements should be approximately correct for the effects with large predictions. The effects with small predictions are unlikely to have approximately normal marginal predictive distributions. This may not cause concern as these effects are unlikely to be *important*.

10.8.1 *Example of Probability Calculation*

The relative performance of the two methods to calculate the required inferential probabilities for the LASSO random effects model are demonstrated using a simulation study. The simulation design employed was the same used to investigate the predictive performance of the different types of predictors. However, in this study the probability $\mathrm{P}(\beta_i > 0|\boldsymbol{y})$ was calculated using both the stochastic and normal approximation methods. The resulting probabilities for all effects for all simulations are plotted in Figure 10.5. This is a plot of 7500 points (500 simulations each with $q = 15$ effects).

In general there is relatively good agreement between the two methods (left hand panel of Figure 10.5). However, closer inspection reveals that near the critical values ($\mathrm{P}(\beta_i > 0|\boldsymbol{y}) = 0.05$ and $\mathrm{P}(\beta_i > 0|\boldsymbol{y}) = 0.95$) the approximate probabilities are higher than the corresponding stochastic ones (right hand panels). This implies that inference using the approximation is likely to be overly aggressive. It could still provide a useful working tool during analysis but this simulation shows that it should not be used formally.

10.9 Re-Analysis of the Prostate Cancer Data

The prostate data was presented in Stamey et al. (1989), analysed in Tibshirani (1996), Osborne et al. (2000) and Chapter 7. A re-analysis highlights the methods presented in this chapter.

The dispersion parameters were estimated using the likelihood from Section 10.3. The estimation process took 40 seconds to converge. This is considerably quicker than the bootstrap method from Chapter 7, which took 14.11 minutes to converge. The dispersion parameter estimates from the two methods are similar, $\hat{\sigma}^2 = 0.504$ and $2\hat{\phi}^2 = 0.073$ for the bootstrap adjustment method and $\hat{\sigma}^2 = 0.494$ and $2\hat{\phi}^2 = 0.087$ for the methods in this chapter.

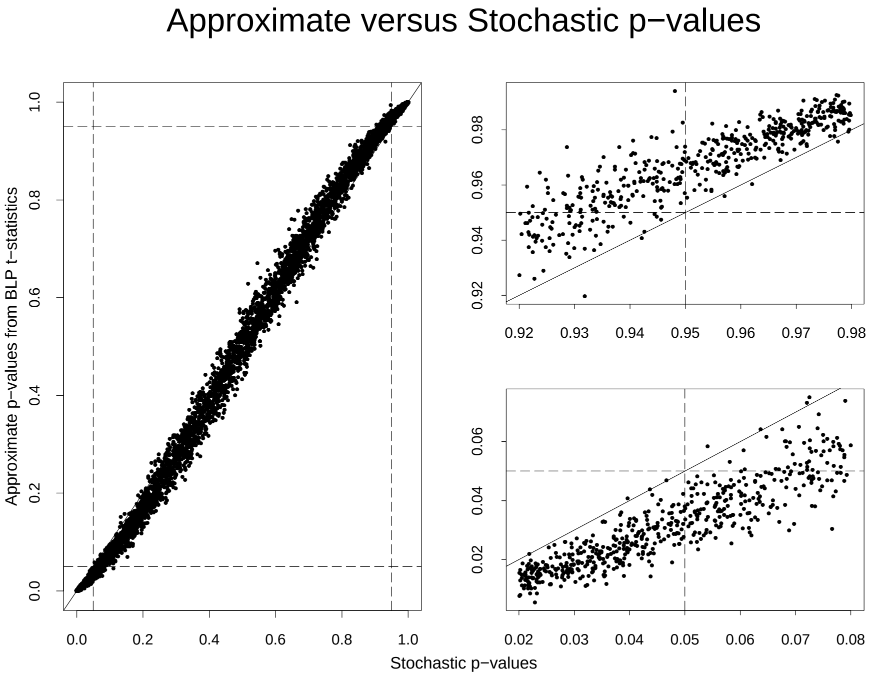

Figure 10.5: Approximate versus stochastic $\mathrm{P}(\beta_i > 0|\boldsymbol{y})$. Left graph is all probabilities, upper right graph is only large probabilities and lower right are only small probabilities. Solid lines are $y = x$. Dashed lines are critical values ($p = 0.05$ and $p = 0.95$).

The LASSO random effects were estimated using the standard LASSO prediction (Tibshirani, 1996 and Chapter 4) along with the prediction methods in Section 10.7. These are presented in Table 10.3. There was generally good agreement between the predictors. The LASSO and conditional best predictor, $\mathrm{CBP}_{\bar{\boldsymbol{\delta}}}$, are the only predictors that calculated zero predictions. $\mathrm{CBP}_{\bar{\boldsymbol{\delta}}}$ predicted more zeros than the LASSO.

Inference was carried out using the stochastic and approximate probability statements in Section 10.8. Both sets of probabilities are given in Table 10.3. There was reasonable quantitative agreement between the two sets of probabilities. The approximate probabilities were always more aggressive (closer to zero or one). Also, there were qualitative differences for the inference based on the two sets of probabilities. In particular, the explanatory variable 'svi' could be considered important using the approximate probabilities but not so using the stochastic probabilities. Further, the explanatory variable 'age' approached significance using the approximate probabilities. This inference agrees well with that from using the simulation hypothesis test from Chapter 7 presented in Table 7.7.

10.10 Summary

In this Chapter, an alternative random effects model specification for the LASSO was presented. The alternative specification leads to an approximate marginal likelihood, whose score equations do not require adjustment. A number of different predictors for the LASSO model were defined and assessed via simulation. The standard LASSO predictor (Tibshirani, 1996) performed adequately and had the appeal of producing identically zero predictions. Inference concerning the LASSO effects was based on the probability statements $\mathrm{P}(\beta_i > 0|\boldsymbol{y})$ or $\mathrm{P}(\beta_i < 0|\boldsymbol{y})$, which can be calculated using importance sampling.

Table 10.3: Re-analysis of the prostate cancer data. $\mathrm{P}_s(\beta_i > 0|\boldsymbol{y})$ and $\mathrm{P}_a(\beta_i > 0|\boldsymbol{y})$ are the stochastic and approximate probabilities described in Section 10.8. Results are directly comparable with Table 7.7.

	Predictions					Probabilities	
	BP	LASSO	BLP	CBP$_{\bar{\boldsymbol{\delta}}}$	CBP$_{\bar{\boldsymbol{\alpha}}}$	$\mathrm{P}_s(\beta_i > 0\|\boldsymbol{y})$	$\mathrm{P}_a(\beta_i > 0\|\boldsymbol{y})$
lcavol	0.634	0.627	0.617	0.636	0.646	1.000	1.000
lweight	0.207	0.202	0.221	0.182	0.215	0.995	0.997
age	-0.091	-0.076	-0.119	0.000	-0.115	0.111	0.066
svi	0.123	0.120	0.141	0.083	0.138	0.946	0.961
lbph	0.266	0.253	0.291	0.256	0.284	0.997	0.999
lcp	-0.044	0.000	-0.067	0.000	-0.074	0.340	0.276
gleason	0.034	0.007	0.043	0.000	0.037	0.655	0.662
pgg45	0.084	0.071	0.101	0.017	0.097	0.808	0.820

Chapter 11

An Alternative LLMM

An alternative specification for the LASSO random effects model was presented in Chapter 10. This alternative model provides a more direct method for estimating the dispersion parameters than the random effects model in Chapter 7. In particular, the need to adjust the score equations using a bootstrap estimate of the expected score was removed. This reduced the amount of computation significantly. However, the random effects model is not flexible enough to handle QTL experimental data. These data require models containing fixed and random normal effects. The alternative random LASSO model is now extended by including LASSO random effects into a normal mixed model. This parallels the extension of the random LASSO model in Chapter 7 to the LASSO linear mixed model (LLMM) in Chapter 8. The alternative LLMM is

$$\boldsymbol{y} = \boldsymbol{X\tau} + \boldsymbol{Zu} + \boldsymbol{L\beta} + \boldsymbol{e} \tag{11.0.1}$$

where

$$\begin{aligned} \boldsymbol{u} &\sim \mathrm{N}_r(\boldsymbol{0}, \sigma^2 \boldsymbol{G}(\boldsymbol{\gamma})), \\ \boldsymbol{e} &\sim \mathrm{N}_n(\boldsymbol{0}, \sigma^2 \boldsymbol{R}(\boldsymbol{\theta})), \\ \beta_i | \delta_i &\sim \mathrm{N}(0, \delta_i) \qquad i = 1, 2, \ldots, q, \\ \delta_i &\sim \exp(2\phi^2), \end{aligned}$$

$\boldsymbol{X}$, $\boldsymbol{Z}$ and $\boldsymbol{L}$ are design matrices for the p fixed effects, the r random normal effects and the q random LASSO effects respectively, σ^2 is the residual variance, $\boldsymbol{G}(\boldsymbol{\gamma})$ is the correlation structure associated with the random normal effects, $\boldsymbol{R}(\boldsymbol{\theta})$ is the correlation structure associated with the residuals, $2\phi^2$ is the variance of the LASSO random effects and $\exp(\delta_i)$ is the exponential distribution with mean δ_i. The vectors of parameters $\boldsymbol{\gamma}$ and $\boldsymbol{\theta}$ completely specify the matrices $\boldsymbol{G}$ and $\boldsymbol{R}$ respectively. Throughout this chapter, no distinction is made between the parameters $\boldsymbol{\gamma}$ and $\boldsymbol{\theta}$. To avoid confusion, define $\boldsymbol{\kappa} = (\boldsymbol{\gamma}^T, \boldsymbol{\theta}^T)^T$.

It is proposed that estimation of the dispersion parameters can be performed using an approximate restricted likelihood based on a partial Laplace approximation (Chapter 3). The approximation and its score equations are presented in this chapter.

In addition some methods for computation, prediction and inference are discussed.

The methods in this chapter have not been implemented. It is expected that they will behave well but it can not be guaranteed.

11.1 Approximate Restricted Likelihood

As in Chapter 10, the distribution of the observed outcomes $\boldsymbol{y}$ conditional on the random variances $\boldsymbol{\delta}$ is normal. In particular,

$$\begin{aligned}\boldsymbol{y}|\boldsymbol{\delta} &\sim \mathrm{N}(\boldsymbol{X\tau}, \sigma^2\boldsymbol{ZGZ}^T + \boldsymbol{LDL}^T + \sigma^2\boldsymbol{R})\\ &= \mathrm{N}(\boldsymbol{X\tau}, \sigma^2\boldsymbol{H}_m + \boldsymbol{LDL}^T)\\ &= \mathrm{N}(\boldsymbol{X\tau}, \boldsymbol{H})\end{aligned}$$

where $\boldsymbol{D} = \operatorname{diagv}(\boldsymbol{\delta})$ and $\boldsymbol{H}_m = \sigma^2\boldsymbol{R} + \sigma^2\boldsymbol{ZGZ}^T$, the variance of the data conditional on the LASSO effects.

Theorem 11.1. *The approximate restricted log-likelihood based on a partial Laplace approximation of the marginal distribution is*

$$\begin{aligned}\ell_r(\sigma^2, 2\phi^2, \boldsymbol{\kappa}; \boldsymbol{y}) = -q\log 2\phi^2 - \frac{1}{2}\log|\bar{\boldsymbol{\Lambda}}| - \frac{1}{2}\log|\bar{\boldsymbol{H}}| - \frac{1}{2}\log|\boldsymbol{X}^T\bar{\boldsymbol{H}}^{-1}\boldsymbol{X}|\\ - \frac{1}{2}\boldsymbol{y}^T\bar{\boldsymbol{P}}\boldsymbol{y} - \frac{1}{2\phi^2}\sum_{i=1}^{q}\exp(\bar{\alpha}_i) + \sum_{i=1}^{q}\bar{\alpha}_i\end{aligned}$$

where $\bar{\boldsymbol{\alpha}}$ is the maximiser of

$$Q(\boldsymbol{\alpha}) = -\frac{1}{2}\log|\boldsymbol{H}| - \frac{1}{2}|\boldsymbol{X}^T\boldsymbol{H}^{-1}\boldsymbol{X}| - \frac{1}{2}\boldsymbol{y}^T\boldsymbol{P}\boldsymbol{y} - \frac{1}{2\phi^2}\sum_{i=1}^{q}\exp(\alpha_i) + \sum_{i=1}^{q}\alpha_i,$$

$\boldsymbol{\Lambda}$ is the matrix of second derivatives of $Q(\boldsymbol{\alpha})$ with respect to $\boldsymbol{\alpha}$,

$$\boldsymbol{P} = \boldsymbol{H}^{-1} - \boldsymbol{H}^{-1}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{H}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{H}^{-1}$$

and the bar notation signifies that the quantity is evaluated at $\boldsymbol{\alpha} = \bar{\boldsymbol{\alpha}}$.

Proof. Following Verbyla (1990) the matrix $\boldsymbol{W} = (\boldsymbol{W}_1, \boldsymbol{W}_2)$ is chosen so that

$$\boldsymbol{W}_1^T\boldsymbol{X} = \boldsymbol{I}_p \qquad \text{and} \qquad \boldsymbol{W}_2^T\boldsymbol{X} = \boldsymbol{0}.$$

Define $\boldsymbol{y}_1 = \boldsymbol{W}_1^T\boldsymbol{y}$ and $\boldsymbol{y}_2 = \boldsymbol{W}_2^T\boldsymbol{y}$. Then the conditional normal distribution can be specified as

$$\begin{pmatrix}\boldsymbol{y}_1\\ \boldsymbol{y}_2\end{pmatrix}\Bigg|\,\boldsymbol{\delta} \sim \mathrm{N}\left(\begin{pmatrix}\boldsymbol{\tau}\\ \boldsymbol{0}\end{pmatrix}, \begin{pmatrix}\boldsymbol{W}_1^T\boldsymbol{HW}_1 & \boldsymbol{W}_1^T\boldsymbol{HW}_2\\ \boldsymbol{W}_2^T\boldsymbol{HW}_1 & \boldsymbol{W}_2^T\boldsymbol{HW}_2\end{pmatrix}\right).$$

Only the joint distribution of $\boldsymbol{y}_2$ and $\boldsymbol{\delta}$ needs to be considered for the partial Laplace approximation. The joint distribution is

$$\begin{aligned} f(\boldsymbol{y}_2|\boldsymbol{\delta})f(\boldsymbol{\delta}) &= \frac{1}{(2\pi)^{\frac{n-p}{2}}(2\phi^2)^q}|\boldsymbol{W}_2^T\boldsymbol{H}\boldsymbol{W}_2|^{-\frac{1}{2}}\exp\left(-\frac{1}{2}\boldsymbol{y}_2^T(\boldsymbol{W}_2^T\boldsymbol{H}\boldsymbol{W}_2)^{-1}\boldsymbol{y}_2 - \frac{1}{2\phi^2}\sum_{i=1}^q \delta_i\right) \\ &\propto \frac{1}{(2\phi^2)^q}\exp\left(-\frac{1}{2}\log|\boldsymbol{H}| - \frac{1}{2}\log|\boldsymbol{X}^T\boldsymbol{H}^{-1}\boldsymbol{X}| - \frac{1}{2}\boldsymbol{y}^T\boldsymbol{P}\boldsymbol{y} - \frac{1}{2\phi^2}\sum_{i=1}^q \delta_i\right). \end{aligned}$$

The restricted marginal likelihood is obtained by integrating this function over $\mathbb{R}_+^q$. The partial Laplace method approximates the integrand by a normal distribution. To make this approximation more accurate a transformation is made, namely $\boldsymbol{\alpha} = \log\boldsymbol{\delta}$. The support of $\boldsymbol{\alpha}$ is the same as that of the approximating normal's support. The joint distribution of $\boldsymbol{y}_2$ and $\boldsymbol{\alpha}$ is

$$\begin{aligned} &f(\boldsymbol{y}_2|\boldsymbol{\alpha})f(\boldsymbol{\alpha}) \\ &\propto \frac{1}{(2\phi^2)^q}\exp\left(-\frac{1}{2}\log|\boldsymbol{H}| - \frac{1}{2}\log|\boldsymbol{X}^T\boldsymbol{H}^{-1}\boldsymbol{X}| - \frac{1}{2}\boldsymbol{y}^T\boldsymbol{P}\boldsymbol{y} - \frac{1}{2\phi^2}\sum_{i=1}^q \exp\alpha_i + \sum_{i=1}^q \alpha_i\right) \\ &= \frac{1}{(2\phi^2)^q}\exp\left(Q(\boldsymbol{\alpha})\right) \qquad\qquad \text{say} \end{aligned}$$

where all quantities are parameterised in terms of $\boldsymbol{\alpha}$ and not $\boldsymbol{\delta}$. The term $\sum_{i=1}^q \alpha_i$ in the exponent arises from the Jacobian of the transformation. The approximate restricted likelihood is based on the distribution

$$\begin{aligned} f(\boldsymbol{y}_2) &= \int_{\mathbb{R}^q} f(\boldsymbol{y}_2|\boldsymbol{\alpha})f(\boldsymbol{\alpha})\partial\boldsymbol{\alpha} \\ &\propto \frac{1}{(2\phi^2)^q}\int_{\mathbb{R}^q}\exp\left(Q(\boldsymbol{\alpha})\right)\partial\boldsymbol{\alpha}. \end{aligned}$$

The first two derivatives of $Q(\boldsymbol{\alpha})$ with respect to $\boldsymbol{\alpha}$ are required for the approximation. Define $\boldsymbol{M} = \boldsymbol{L}^T\boldsymbol{P}\boldsymbol{L}$ and $\boldsymbol{w} = \boldsymbol{L}^T\boldsymbol{P}\boldsymbol{y}$. Consider first the derivatives with respect to individual elements of $\boldsymbol{\alpha}$, α_i say.

$$\begin{aligned} \frac{\partial Q(\boldsymbol{\alpha})}{\partial\alpha_i} &= -\frac{1}{2}\exp(\alpha_i)\,\mathrm{tr}\left(\boldsymbol{H}^{-1}\boldsymbol{l}_i\boldsymbol{l}_i^T\right) - \frac{1}{2}\exp(\alpha_i)\,\mathrm{tr}\left((\boldsymbol{X}^T\boldsymbol{H}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{H}^{-1}\boldsymbol{l}_i\boldsymbol{l}_i^T\boldsymbol{H}^{-1}\boldsymbol{X}\right) \\ &\qquad + \frac{1}{2}\exp(\alpha_i)\,\boldsymbol{y}^T\boldsymbol{P}\boldsymbol{l}_i\boldsymbol{l}_i^T\boldsymbol{P}\boldsymbol{y} - \frac{1}{2\phi^2}\exp(\alpha_i) + 1 \\ &= -\exp(\alpha_i)\left(\frac{1}{2}\boldsymbol{l}_i^T\boldsymbol{P}\boldsymbol{l}_i - \frac{1}{2}w_i^2 + \frac{1}{2\phi^2}\right) + 1 \\ &= -\exp(\alpha_i)\,b_i + 1 \qquad\qquad \text{say.} \end{aligned}$$

The diagonal elements of the Hessian are

$$\frac{\partial^2 Q(\boldsymbol{\alpha})}{\partial\alpha_i^2} = -\exp(\alpha_i)\,b_i + \exp(\alpha_i)\left(\frac{1}{2}M_{ii}^2 - w_i^2 M_{ii}\right)$$

and the off-diagonal elements are $(i \neq j)$

$$\frac{\partial^2 Q(\boldsymbol{\alpha})}{\partial \alpha_i \partial \alpha_j} = \exp(\alpha_i) \exp(\alpha_j) \left(\frac{1}{2} M_{ij}^2 - w_i w_j M_{ij} \right).$$

The first derivative is

$$\begin{aligned}\frac{\partial Q(\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}} &= -\exp(\boldsymbol{\alpha}) \circ \left(\frac{1}{2} \operatorname{vecd}(\boldsymbol{M}) - \frac{1}{2} \boldsymbol{w} \circ \boldsymbol{w} + \frac{1}{2\phi^2} \mathbf{1} \right) + \mathbf{1} \\ &= -\exp(\boldsymbol{\alpha}) \circ \boldsymbol{b} + \mathbf{1}.\end{aligned}$$

The Hessian is

$$\begin{aligned}\frac{\partial^2 Q(\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha} \partial \boldsymbol{\alpha}^T} &= -\operatorname{diagv}(\exp(\boldsymbol{\alpha}) \circ \boldsymbol{b}) + \exp(\boldsymbol{\alpha}) \exp(\boldsymbol{\alpha})^T \circ \left(\frac{1}{2} \boldsymbol{M} \circ \boldsymbol{M} - \boldsymbol{w}\boldsymbol{w}^T \circ \boldsymbol{M} \right) \\ &= -\boldsymbol{\Lambda} \qquad \text{say.}\end{aligned}$$

The distribution $f(\boldsymbol{y}_2)$ is approximately

$$\begin{aligned}f(\boldsymbol{y}_2) &\propto \frac{1}{(2\phi^2)^q} \exp(Q(\bar{\boldsymbol{\alpha}})) |\bar{\boldsymbol{\Lambda}}|^{-\frac{1}{2}} \\ &= \frac{1}{(2\phi^2)^q} |\bar{\boldsymbol{\Lambda}}|^{-\frac{1}{2}} |\bar{\boldsymbol{H}}|^{-\frac{1}{2}} |\boldsymbol{X}^T \bar{\boldsymbol{H}}^{-1} \boldsymbol{X}|^{-\frac{1}{2}} \exp\left(-\frac{1}{2} \boldsymbol{y}^T \bar{\boldsymbol{P}} \boldsymbol{y} - \frac{1}{2\phi^2} \sum_{i=1}^{q} \exp(\bar{\alpha}_i) + \sum_{i=1}^{q} \bar{\alpha}_i \right)\end{aligned}$$

and the approximate log-likelihood is obtained by taking the log of this expression. □

The matrix $\boldsymbol{\Lambda}$ is the observed information for $\boldsymbol{\alpha}$. It is not guaranteed to be positive definite, which may cause problems when used in estimation. Its expectation, conditional on $\boldsymbol{\alpha}$, is positive definite and can be used in calculating $\bar{\boldsymbol{\alpha}}$. In Chapter 10 however, the use of the expected information matrix for $\boldsymbol{\alpha}$ in the likelihood approximation did not produce satisfactory dispersion parameter estimates. It is expected that the same result applies for the alternative LLMM.

Theorem 11.2. *The conditional expectation for* $\boldsymbol{\alpha}$ *is*

$$\boldsymbol{\Lambda}_e = \operatorname{diagv}\left(\frac{1}{2\phi^2} \exp(\boldsymbol{\alpha}) \right) + \exp(\boldsymbol{\alpha}) \exp(\boldsymbol{\alpha})^T \circ \boldsymbol{M} \circ \boldsymbol{M}$$

where $\boldsymbol{M} = \boldsymbol{L}^T \boldsymbol{P} \boldsymbol{L}$.

Proof. Define $\boldsymbol{w}$ and $\boldsymbol{M}$ as in the proof of Theorem 11.1. Then

$$\begin{aligned}\operatorname{E}(w_i w_j | \boldsymbol{\alpha}) &= \operatorname{cov}(w_i, w_j | \boldsymbol{\alpha}) + \operatorname{E}(w_i | \boldsymbol{\alpha}) \operatorname{E}(w_j | \boldsymbol{\alpha}) \\ &= \boldsymbol{l}_i^T \boldsymbol{P} \boldsymbol{l}_j \qquad \text{and} \\ \operatorname{E}\left(\boldsymbol{w}\boldsymbol{w}^T | \boldsymbol{\alpha}\right) &= \boldsymbol{L}^T \boldsymbol{P} \boldsymbol{L}.\end{aligned}$$

The conditional expected information for $\boldsymbol{\alpha}$ is

$$\begin{aligned}\operatorname{E}(\boldsymbol{\Lambda} | \boldsymbol{\alpha}) &= \operatorname{diagv}(\exp(\boldsymbol{\alpha}) \circ \operatorname{E}(\boldsymbol{b} | \boldsymbol{\alpha})) + \frac{1}{2} \exp(\boldsymbol{\alpha}) \exp(\boldsymbol{\alpha})^T \circ \boldsymbol{M} \circ \boldsymbol{M} \\ &= \operatorname{diagv}\left(\frac{1}{2\phi^2} \exp(\boldsymbol{\alpha}) \right) + \frac{1}{2} \exp(\boldsymbol{\alpha}) \exp(\boldsymbol{\alpha})^T \circ \boldsymbol{M} \circ \boldsymbol{M}.\end{aligned}$$

□

11.2 Score Equations and Their Components

The score equations require the derivatives of $\bar{\boldsymbol{\alpha}}$ and $\bar{\boldsymbol{\Lambda}}$ with respect to σ^2, $2\phi^2$ and $\{\kappa_k\}$. These are non-trivial results and are presented before the score equations.

Theorem 11.3. *The derivatives of* $\bar{\boldsymbol{\alpha}}$ *are*

$$\begin{aligned}\frac{\partial \bar{\boldsymbol{\alpha}}}{\partial \sigma^2} &= \left(\left(\frac{1}{2}\bar{\boldsymbol{M}} \circ \bar{\boldsymbol{M}} - \bar{\boldsymbol{w}}\bar{\boldsymbol{w}}^T \circ \bar{\boldsymbol{M}}\right)\bar{\boldsymbol{D}} - \bar{\boldsymbol{D}}^{-1}\right)^{-1}\left(\bar{\boldsymbol{w}} \circ \bar{\boldsymbol{w}}_2 - \frac{1}{2}\text{vecd}\left(\bar{\boldsymbol{M}}_2\right)\right), \\ \frac{\partial \bar{\boldsymbol{\alpha}}}{\partial 2\phi^2} &= -\frac{1}{4\phi^4}\left(\left(\frac{1}{2}\bar{\boldsymbol{M}} \circ \bar{\boldsymbol{M}} - \bar{\boldsymbol{w}}\bar{\boldsymbol{w}}^T \circ \bar{\boldsymbol{M}}\right)\bar{\boldsymbol{D}} - \bar{\boldsymbol{D}}^{-1}\right)^{-1}\boldsymbol{1} \qquad \textit{and} \\ \frac{\partial \bar{\boldsymbol{\alpha}}}{\partial \kappa_k} &= \sigma^2\left(\left(\frac{1}{2}\bar{\boldsymbol{M}} \circ \bar{\boldsymbol{M}} - \bar{\boldsymbol{w}}\bar{\boldsymbol{w}}^T \circ \bar{\boldsymbol{M}}\right)\bar{\boldsymbol{D}} - \bar{\boldsymbol{D}}^{-1}\right)^{-1}\left(\bar{\boldsymbol{w}} \circ \bar{\boldsymbol{w}}_{2k} - \frac{1}{2}\text{vecd}\left(\bar{\boldsymbol{M}}_{2k}\right)\right)\end{aligned}$$

where $\bar{\boldsymbol{M}} = \boldsymbol{L}^T\bar{\boldsymbol{P}}\boldsymbol{L}$, $\bar{\boldsymbol{w}} = \boldsymbol{L}^T\bar{\boldsymbol{P}}\boldsymbol{y}$, $\bar{\boldsymbol{M}}_2 = \boldsymbol{L}^T\bar{\boldsymbol{P}}\boldsymbol{H}_m\bar{\boldsymbol{P}}\boldsymbol{L}$, $\bar{\boldsymbol{w}}_2 = \boldsymbol{L}^T\bar{\boldsymbol{P}}\boldsymbol{H}_m\bar{\boldsymbol{P}}\boldsymbol{y}$, $\bar{\boldsymbol{M}}_{2k} = \boldsymbol{L}^T\bar{\boldsymbol{P}}\dot{\boldsymbol{H}}_{mk}\bar{\boldsymbol{P}}\boldsymbol{L}$, $\bar{\boldsymbol{w}}_{2k} = \boldsymbol{L}^T\bar{\boldsymbol{P}}\dot{\boldsymbol{H}}_{mk}\bar{\boldsymbol{P}}\boldsymbol{y}$ *and* $\dot{\boldsymbol{H}}_{mk}$ *is the derivative of* $\boldsymbol{H}_m$ *with respect to the* k^{th} *dispersion parameter* κ_k.

Proof. Only the proof for $\partial\bar{\boldsymbol{\alpha}}/\partial\kappa_k$ will be presented as the others follow similarly. They are extensions of the results in Theorem 10.3.

The estimating equation for the i^{th} random variance parameter α_i, $\partial Q(\boldsymbol{\alpha})/\partial\alpha_i = 0$, implies that

$$\begin{aligned} 0 &= -\frac{1}{2}\boldsymbol{l}_i^T\bar{\boldsymbol{P}}\boldsymbol{l}_i + \frac{1}{2}\bar{w}_i^2 - \frac{1}{2\phi^2} + \exp\left(-\bar{\alpha}_i\right) \\ \therefore \qquad 0 &= \frac{1}{2}\boldsymbol{l}_i^T\bar{\boldsymbol{P}}\left(\sigma^2\dot{\boldsymbol{H}}_{mk} + \boldsymbol{L}\bar{\boldsymbol{D}}_{\kappa_k}\boldsymbol{L}^T\right)\bar{\boldsymbol{P}}\boldsymbol{l}_i - \bar{w}_i\boldsymbol{l}_i\bar{\boldsymbol{P}}\left(\sigma^2\dot{\boldsymbol{H}}_{mk} + \boldsymbol{L}\bar{\boldsymbol{D}}_{\kappa_k}\boldsymbol{L}^T\right)\bar{\boldsymbol{P}}\boldsymbol{y} \\ &\qquad - \exp\left(-\bar{\alpha}_i\right)\frac{\partial\bar{\alpha}_i}{\partial\kappa_k}\end{aligned}$$

as

$$\begin{aligned}\frac{\partial\bar{\boldsymbol{P}}}{\partial\kappa_k} &= -\bar{\boldsymbol{P}}\left(\sigma^2\dot{\boldsymbol{H}}_{mk} + \boldsymbol{L}\bar{\boldsymbol{D}}_{\kappa_k}\boldsymbol{L}^T\right)\bar{\boldsymbol{P}} \qquad \text{and} \\ \bar{\boldsymbol{D}}_{\kappa_k} &= \frac{\partial\bar{\boldsymbol{D}}}{\partial\kappa_k} \\ &= \text{diagv}\left(\exp\left(\bar{\boldsymbol{\alpha}}\right) \circ \frac{\partial\bar{\boldsymbol{\alpha}}}{\partial\kappa_k}\right).\end{aligned}$$

Continuing with the implicit differentiation for the i^{th} random variance parameter yields

$$\begin{aligned}&\sigma^2\left(\bar{w}_i\boldsymbol{l}_i^T\bar{\boldsymbol{P}}\dot{\boldsymbol{H}}_{mk}\bar{\boldsymbol{P}}\boldsymbol{y} - \frac{1}{2}\boldsymbol{l}_i^T\bar{\boldsymbol{P}}\dot{\boldsymbol{H}}_{mk}\bar{\boldsymbol{P}}\boldsymbol{l}_i\right) \\ &\qquad = \frac{1}{2}\boldsymbol{l}_i^T\bar{\boldsymbol{P}}\boldsymbol{L}\bar{\boldsymbol{D}}_{\kappa_k}\boldsymbol{L}^T\bar{\boldsymbol{P}}\boldsymbol{l}_i - \bar{w}_i\boldsymbol{l}_i^T\bar{\boldsymbol{P}}\boldsymbol{L}\bar{\boldsymbol{D}}_{\kappa_k}\boldsymbol{L}^T\bar{\boldsymbol{P}}\boldsymbol{y} - \exp\left(-\bar{\alpha}_i\right)\frac{\partial\bar{\alpha}_i}{\partial\kappa_k} \\ &\qquad = \sum_{j=1}^{q}\left(\left(\frac{1}{2}\bar{M}_{ij}^2 - \bar{w}_i\bar{w}_j\bar{M}_{ij}\right)\exp\left(\bar{\alpha}_j\right)\frac{\partial\bar{\alpha}_j}{\partial\kappa_k}\right) - \exp\left(-\bar{\alpha}_i\right)\frac{\partial\bar{\alpha}_i}{\partial\kappa_k} \\ &\qquad = \left(\frac{1}{2}\bar{\boldsymbol{M}}_{i\cdot} \circ \bar{\boldsymbol{M}}_{i\cdot} - \bar{w}_i\bar{\boldsymbol{w}}^T \circ \bar{\boldsymbol{M}}_{i\cdot}\right)\left(\exp\left(\bar{\boldsymbol{\alpha}}\right)\frac{\partial\bar{\boldsymbol{\alpha}}}{\partial\kappa_k}\right) - \exp\left(-\bar{\alpha}_i\right)\frac{\partial\bar{\alpha}_i}{\partial\kappa_k}\end{aligned}$$

where $\bar{\boldsymbol{M}}_{i\cdot}$ is the i^{th} row of $\bar{\boldsymbol{M}}$. The expression for all the random dispersion parameters is

$$\sigma^2\left(\bar{\boldsymbol{w}}\boldsymbol{L}^T\bar{\boldsymbol{P}}\dot{\boldsymbol{H}}_{mk}\bar{\boldsymbol{P}}\boldsymbol{y} - \text{vecd}\left(\frac{1}{2}\boldsymbol{L}^T\bar{\boldsymbol{P}}\dot{\boldsymbol{H}}_{mk}\bar{\boldsymbol{P}}\boldsymbol{L}\right)\right) = \left(\left(\frac{1}{2}\bar{\boldsymbol{M}}\circ\bar{\boldsymbol{M}} - \bar{\boldsymbol{w}}\bar{\boldsymbol{w}}^T\circ\bar{\boldsymbol{M}}\right)\bar{\boldsymbol{D}} - \bar{\boldsymbol{D}}^{-1}\right)\frac{\partial\bar{\boldsymbol{\alpha}}}{\partial\kappa_k}.$$

Solving for the derivative completes the proof. □

Theorem 11.4. *The derivatives of the observed information for $\boldsymbol{\alpha}$, evaluated at $\bar{\boldsymbol{\alpha}}$ are*

$$\begin{aligned}
\frac{\partial\bar{\boldsymbol{\Lambda}}}{\partial\sigma^2} &= \bar{\boldsymbol{D}}_{\sigma^2}\text{diagv}\left(\frac{1}{2}\text{vecd}\left(\bar{\boldsymbol{M}}\right) - \frac{1}{2}\bar{\boldsymbol{w}}\circ\bar{\boldsymbol{w}} + \frac{1}{2\phi^2}\mathbf{1}\right)\\
&+ \text{diagv}\left(\exp\left(\bar{\boldsymbol{\alpha}}\right)\circ\left(-\frac{1}{2}\text{vecd}\left(\bar{\boldsymbol{M}}_2 - \bar{\boldsymbol{M}}\bar{\boldsymbol{D}}_{\sigma^2}\bar{\boldsymbol{M}}\right) + \bar{\boldsymbol{w}}\circ\bar{\boldsymbol{w}}_2 + \bar{\boldsymbol{w}}\circ\bar{\boldsymbol{M}}\bar{\boldsymbol{D}}_{\sigma^2}\bar{\boldsymbol{w}}\right)\right)\\
&- \left(\bar{\boldsymbol{\Delta}}_{\sigma^2} + \bar{\boldsymbol{\Delta}}_{\sigma^2}^T\right)\circ\left(\frac{1}{2}\bar{\boldsymbol{M}}\circ\bar{\boldsymbol{M}} - \bar{\boldsymbol{w}}\bar{\boldsymbol{w}}^T\circ\bar{\boldsymbol{M}}\right)\\
&- \exp\left(\bar{\boldsymbol{\alpha}}\right)\exp\left(\bar{\boldsymbol{\alpha}}\right)^T\circ\left(\left(\bar{\boldsymbol{w}}\bar{\boldsymbol{w}}^T - \bar{\boldsymbol{M}}\right)\circ\bar{\boldsymbol{M}}_2 + \left(\bar{\boldsymbol{w}}\bar{\boldsymbol{w}}_2^T + \bar{\boldsymbol{w}}_2\bar{\boldsymbol{w}}^T\right)\circ\bar{\boldsymbol{M}}\right)\\
&- \exp\left(\bar{\boldsymbol{\alpha}}\right)\exp\left(\bar{\boldsymbol{\alpha}}\right)^T\circ\left(\left(\bar{\boldsymbol{w}}\bar{\boldsymbol{w}}^T - \bar{\boldsymbol{M}}\right)\circ\bar{\boldsymbol{M}}\bar{\boldsymbol{D}}_{\sigma^2}\bar{\boldsymbol{M}} + \left(\left(\bar{\boldsymbol{M}}\bar{\boldsymbol{D}}_{\sigma^2}\bar{\boldsymbol{w}}\right)\bar{\boldsymbol{w}}^T + \bar{\boldsymbol{w}}\left(\bar{\boldsymbol{M}}\bar{\boldsymbol{D}}_{\sigma^2}\bar{\boldsymbol{w}}\right)^T\right)\circ\bar{\boldsymbol{M}}\right),\\
\frac{\partial\bar{\boldsymbol{\Lambda}}}{\partial 2\phi^2} &= \bar{\boldsymbol{D}}_{2\phi^2}\text{diagv}\left(\frac{1}{2}\text{vecd}\left(\bar{\boldsymbol{M}}\right) - \frac{1}{2}\bar{\boldsymbol{w}}\circ\bar{\boldsymbol{w}} + \frac{1}{2\phi^2}\mathbf{1}\right)\\
&+ \text{diagv}\left(\exp\left(\bar{\boldsymbol{\alpha}}\right)\circ\left(-\frac{1}{2}\text{vecd}\left(\bar{\boldsymbol{M}}\bar{\boldsymbol{D}}_{2\phi^2}\bar{\boldsymbol{M}}\right) + \bar{\boldsymbol{w}}\circ\bar{\boldsymbol{M}}\bar{\boldsymbol{D}}_{2\phi^2}\bar{\boldsymbol{w}} - \frac{1}{4\phi^4}\mathbf{1}\right)\right)\\
&- \left(\bar{\boldsymbol{\Delta}}_{2\phi^2} + \bar{\boldsymbol{\Delta}}_{2\phi^2}^T\right)\circ\left(\frac{1}{2}\bar{\boldsymbol{M}}\circ\bar{\boldsymbol{M}} - \bar{\boldsymbol{w}}\bar{\boldsymbol{w}}^T\circ\bar{\boldsymbol{M}}\right)\\
&- \exp\left(\bar{\boldsymbol{\alpha}}\right)\exp\left(\bar{\boldsymbol{\alpha}}\right)^T\circ\left(\left(\bar{\boldsymbol{w}}\bar{\boldsymbol{w}}^T - \bar{\boldsymbol{M}}\right)\circ\bar{\boldsymbol{M}}\bar{\boldsymbol{D}}_{2\phi^2}\bar{\boldsymbol{M}} + \left(\left(\bar{\boldsymbol{M}}\bar{\boldsymbol{D}}_{2\phi^2}\bar{\boldsymbol{w}}\right)\bar{\boldsymbol{w}}^T + \bar{\boldsymbol{w}}\left(\bar{\boldsymbol{M}}\bar{\boldsymbol{D}}_{2\phi^2}\bar{\boldsymbol{w}}\right)^T\right)\circ\bar{\boldsymbol{M}}\right)
\end{aligned}$$

and

$$\begin{aligned}
\frac{\partial\bar{\boldsymbol{\Lambda}}}{\partial\kappa_k} &= \bar{\boldsymbol{D}}_{\kappa_k}\text{diagv}\left(\frac{1}{2}\text{vecd}\left(\bar{\boldsymbol{M}}\right) - \frac{1}{2}\bar{\boldsymbol{w}}\circ\bar{\boldsymbol{w}} + \frac{1}{2\phi^2}\mathbf{1}\right)\\
&+ \text{diagv}\left(\exp\left(\bar{\boldsymbol{\alpha}}\right)\circ\left(-\frac{1}{2}\text{vecd}\left(\sigma^2\dot{\boldsymbol{M}}_{2k} + \bar{\boldsymbol{M}}\bar{\boldsymbol{D}}_{\kappa_k}\bar{\boldsymbol{M}}\right) + \sigma^2\bar{\boldsymbol{w}}\circ\bar{\boldsymbol{w}}_{2k} + \bar{\boldsymbol{w}}\circ\bar{\boldsymbol{M}}\bar{\boldsymbol{D}}_{\kappa_k}\bar{\boldsymbol{M}}\right)\right)\\
&- \left(\bar{\boldsymbol{\Delta}}_{\kappa_k} + \bar{\boldsymbol{\Delta}}_{\kappa_k}^T\right)\circ\left(\frac{1}{2}\bar{\boldsymbol{M}}\circ\bar{\boldsymbol{M}} - \bar{\boldsymbol{w}}\bar{\boldsymbol{w}}^T\circ\bar{\boldsymbol{M}}\right)\\
&- \sigma^2\exp\left(\bar{\boldsymbol{\alpha}}\right)\exp\left(\bar{\boldsymbol{\alpha}}\right)^T\circ\left(\left(\bar{\boldsymbol{w}}\bar{\boldsymbol{w}}^T - \bar{\boldsymbol{M}}\right)\circ\bar{\boldsymbol{M}}_{2k} + \left(\bar{\boldsymbol{w}}\bar{\boldsymbol{w}}_{2k}^T + \bar{\boldsymbol{w}}_{2k}\bar{\boldsymbol{w}}^T\right)\circ\bar{\boldsymbol{M}}\right)\\
&- \exp\left(\bar{\boldsymbol{\alpha}}\right)\exp\left(\bar{\boldsymbol{\alpha}}\right)^T\circ\left(\left(\bar{\boldsymbol{w}}\bar{\boldsymbol{w}}^T - \bar{\boldsymbol{M}}\right)\circ\bar{\boldsymbol{M}}\bar{\boldsymbol{D}}_{\kappa_k}\bar{\boldsymbol{M}} + \left(\left(\bar{\boldsymbol{M}}\bar{\boldsymbol{D}}_{\kappa_k}\bar{\boldsymbol{w}}\right)\bar{\boldsymbol{w}}^T + \bar{\boldsymbol{w}}\left(\bar{\boldsymbol{M}}\bar{\boldsymbol{D}}_{\kappa_k}\bar{\boldsymbol{w}}\right)^T\right)\circ\bar{\boldsymbol{M}}\right)
\end{aligned}$$

where the matrices $\bar{\boldsymbol{\Delta}}_{\sigma^2}$, $\bar{\boldsymbol{\Delta}}_{2\phi^2}$ and $\bar{\boldsymbol{\Delta}}_{\kappa_k}$ are all similarly defined. For example

$$\bar{\boldsymbol{\Delta}}_{\sigma^2} = \left(\exp\left(\bar{\boldsymbol{\alpha}}\right)\circ\frac{\partial\bar{\boldsymbol{\alpha}}}{\partial\sigma^2}\right)\exp\left(\bar{\boldsymbol{\alpha}}\right)^T.$$

Proof. The proof closely follows that of Theorem 10.4. It is omitted. □

Theorem 11.5. *The score equations for σ^2, $2\phi^2$ and κ_k are*

$$
\begin{aligned}
\frac{\partial \ell_r}{\partial \sigma^2} &= -\frac{1}{2}\mathrm{tr}\left(\bar{\boldsymbol{\Lambda}}^{-1}\frac{\partial \bar{\boldsymbol{\Lambda}}}{\partial \sigma^2}\right) - \frac{1}{2}\mathrm{tr}\left(\bar{\boldsymbol{P}}\left(\boldsymbol{H}_m + \boldsymbol{L}\bar{\boldsymbol{D}}_{\sigma^2}\boldsymbol{L}^T\right)\right) \\
&\quad + \frac{1}{2}\boldsymbol{y}^T\bar{\boldsymbol{P}}(\boldsymbol{H}_m + \boldsymbol{L}\bar{\boldsymbol{D}}_{\sigma^2}\boldsymbol{L}^T)\bar{\boldsymbol{P}}\boldsymbol{y} - \frac{1}{2\phi^2}\sum_{i=1}^{q}\exp\left(\bar{\alpha}_i\right)\frac{\partial \alpha_i}{\partial \sigma^2} + \sum_{i=1}^{q}\frac{\partial \bar{\alpha}_i}{\partial \sigma^2}, \\
\frac{\partial \ell_r}{\partial 2\phi^2} &= -\frac{q}{2\phi^2} - \frac{1}{2}\mathrm{tr}\left(\bar{\boldsymbol{\Lambda}}^{-1}\frac{\partial \bar{\boldsymbol{\Lambda}}}{\partial 2\phi^2}\right) - \frac{1}{2}\mathrm{tr}\left(\boldsymbol{L}^T\bar{\boldsymbol{P}}\boldsymbol{L}\bar{\boldsymbol{D}}_{2\phi^2}\right) + \frac{1}{2}\boldsymbol{y}^T\bar{\boldsymbol{P}}\boldsymbol{L}\bar{\boldsymbol{D}}_{2\phi^2}\boldsymbol{L}^T\bar{\boldsymbol{P}}\boldsymbol{y} \\
&\quad + \frac{1}{4\phi^4}\sum_{i=1}^{q}\exp\left(\bar{\alpha}_i\right) - \frac{1}{2\phi^2}\sum_{i=1}^{q}\exp\left(\bar{\alpha}_i\right)\frac{\partial \bar{\alpha}_i}{\partial 2\phi^2} + \sum_{i=1}^{q}\frac{\partial \bar{\alpha}_i}{\partial 2\phi^2}
\end{aligned}
$$

and

$$
\begin{aligned}
\frac{\partial \ell_r}{\partial \kappa_k} &= -\frac{1}{2}\mathrm{tr}\left(\bar{\boldsymbol{\Lambda}}^{-1}\frac{\partial \bar{\boldsymbol{\Lambda}}}{\partial \kappa_k}\right) - \frac{1}{2}\mathrm{tr}\left(\bar{\boldsymbol{P}}\left(\sigma^2\dot{\boldsymbol{H}}_{mk} + \boldsymbol{L}\bar{\boldsymbol{D}}_{\kappa_k}\boldsymbol{L}^T\right)\right) \\
&\quad + \frac{1}{2}\boldsymbol{y}^T\bar{\boldsymbol{P}}\left(\sigma^2\dot{\boldsymbol{H}}_{mk} + \boldsymbol{L}\bar{\boldsymbol{D}}_{\kappa_k}\boldsymbol{L}^T\right)\bar{\boldsymbol{P}}\boldsymbol{y} - \frac{1}{2\phi^2}\sum_{i=1}^{q}\exp\left(\bar{\alpha}_i\right)\frac{\partial \bar{\alpha}_i}{\partial \kappa_k} + \sum_{i=1}^{q}\frac{\partial \bar{\alpha}_i}{\partial \kappa_k}
\end{aligned}
$$

where all quantities are given in Theorems 11.3 and 11.4.

Proof. The score equations are directly obtained using the results in Appendix B. □

11.3 Computation

11.3.1 *Estimating the Random Dispersion Effects*

The estimating equation for $\boldsymbol{\alpha}$ in the alternative LLMM, namely $\partial Q(\boldsymbol{\alpha})/\partial \boldsymbol{\alpha}$, is a slight extension of the corresponding estimating equation for the random LASSO model. In particular, instances of the matrix $\boldsymbol{H}$ are replaced by $\boldsymbol{P}$. Using this observation, it is expected that the same estimating approach could be used to estimate $\bar{\boldsymbol{\alpha}}$ in the alternative LLMM.

11.3.2 *Estimation of Dispersion Parameters*

An extension of the algorithm used to calculate the random LASSO model in Chapter 10 could be employed to estimate the dispersion parameters. This would require calculating a numerical approximation to the Hessian of the dispersion parameters each time they are updated. If there are only a few dispersion parameters then this should work well. That is, the algorithm should be stable and relatively inexpensive. However, increasing the number of dispersion parameters increases the number of times the matrix $\bar{\boldsymbol{P}}$ is evaluated. The computational cost of these calculations may be prohibitive. Also, as the number of dispersion parameters increases the numerical Hessian may become less stable and may even be singular. This could cause the algorithm to fail.

These computational problems were not experienced in the estimation method described for the LLMM in Chapter 8. In that chapter, an approximate information based on the average information (Chapter 3) was employed. The average information matrix requires only the calculation of quadratic forms, available from absorption of suitably chosen working variates. These quadratic forms also appear in the score equations and hence can be re-used.

The quadratic forms in the score equations for the alternative LLMM (Theorem 11.5) could be calculated using suitably chosen working variates. However, neither the observed nor expected information matrices are available and hence it is not clear how to suitably approximate the elements of the information with quadratic forms. Further, different quadratic forms would be required for the scores and the information as they are both evaluated at $\bar{\boldsymbol{\alpha}}$ or a derivative of it. This would reduce the computational gain made by using only quadratic forms.

11.4 Prediction and Estimation of Effects

Once the dispersion parameters have been estimated the fixed effects can be estimated and both sets of random effects can be predicted using the LASSO mixed model equations (LMME; Chapter 8). This estimation and prediction method calculates the effects at the joint distribution's maximum enabling the LASSO effects to be predicted to be identically zero. These predictions from the LMME will not be best as they are not expected values. However, in Chapter 8 the fixed effect estimates were shown to be approximately unbiased and hence, loss in prediction error (if significant) will mostly arise from prediction of the random effects.

Estimation of the fixed effects could also be conducted through the partial Laplace approximation (Chapter 3). The estimates obtained in this manner may be well approximated by the best linear unbiased estimators (BLUEs), especially if the estimates from the partial Laplace method are approximately linear. The BLUEs are easily calculated as they are the generalised least squares estimates (Searle et al., 1992, Chapter 12). The variance of the outcomes is

$$\begin{aligned} \operatorname{var}(\boldsymbol{y}) &= \sigma^2 \boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}^T + 2\phi^2 \boldsymbol{L}\boldsymbol{L}^T + \sigma^2 \boldsymbol{R} \\ &= \boldsymbol{H}_* \qquad\qquad \text{say} \end{aligned}$$

and the BLUE for the fixed effects is

$$\hat{\boldsymbol{\tau}} = (\boldsymbol{X}^T \boldsymbol{H}_*^{-1} \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{H}_*^{-1} \boldsymbol{y}.$$

Estimating the fixed effects and, predicting the random normal and the random LASSO effects assuming linearity gives the best linear unbiased predictor (BLUP). This is a direct extension of the best linear predictor in Chapter 10. This method gives the BLUE for the fixed effects and the following predictors for the random normal effects and the random LASSO effects

$$\begin{aligned} \tilde{\boldsymbol{u}} &= \sigma^2 \boldsymbol{G}\boldsymbol{Z}^T \boldsymbol{P}_*^{-1} \boldsymbol{y} \qquad\qquad \text{and} \\ \tilde{\boldsymbol{\beta}} &= 2\phi^2 \boldsymbol{L}^T \boldsymbol{P}_*^{-1} \boldsymbol{y} \end{aligned}$$

where $\boldsymbol{P}_* = \boldsymbol{H}_*^{-1} - \boldsymbol{H}_*^{-1}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{H}_*^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}\boldsymbol{H}_*^{-1}$. These predictors are obtained using an extension of Theorem 3.1. The BLUPs will also not be the best predictors as the constraint of linearity is imposed.

Calculation of the best predictors (BP) for the LASSO random effects could proceed using the importance sampling approach introduced in Chapter 10. There, samples were generated using the conditional distribution of $\boldsymbol{\beta}|\boldsymbol{y}$. In the mixed model setting the conditioning should be performed on $\boldsymbol{y}_2$ and not $\boldsymbol{y}$ as the former does not contain the fixed effects - see Cullis et al. (2006, Chapter 6) for a discussion for normal mixed models. A suitable proposal distribution might be a multivariate normal with mean $\tilde{\boldsymbol{\beta}}$ and variance $2\phi^2\boldsymbol{I} - 4\phi^4\boldsymbol{L}^T\boldsymbol{P}_*\boldsymbol{L}$. The random normal and random LASSO effects are assumed to be independent and hence the best predictor for the random normal is the BLUP $\tilde{\boldsymbol{u}}$. The fixed effects could be estimated using the BLUE $\hat{\boldsymbol{\tau}}$.

The conditional best predictors, introduced in Chapter 10, could also be used for the LLMM model. Conditional on the random dispersion effects the model is normal and BLUP should be used. In Chapter 10, the choice of value to condition on was not clear. The same applies for the LLMM.

11.5 Inference for Random Effects

Inference for the random LASSO effects could be performed using the principles given in Chapter 10. An obvious choice for the proposal distribution is the multivariate normal used for calculating the LASSO BP.

Inference for the random normal effects could be performed using the methods for the normal mixed model. This can be done as the random normal effects and the random LASSO effects are assumed independent.

11.6 Summary

In this chapter, the methods for estimation, prediction and inference for the alternative LLMM were outlined. Estimation, is based on an approximate residual likelihood based on the marginal distribution of the data. These methods have not been implemented.

Chapter 12

Closing Remarks

12.1 Overview

The motivation for the methods developed in this thesis is the identification of quantitative trait loci (QTL). The analysis of QTL mapping data can be thought of as a model selection problem as there are many genetic markers (explanatory variables) scored throughout the population. The markers or the intervals between them, form proxies for QTL. Only a (small) number of these markers/intervals will be linked to a QTL and hence affect the quantitative trait of interest. The task of finding the most appropriate subset(s) of markers/intervals is far from trivial as the number of potential models can be very large, for example in just one family of the Davies' Gene Mapping Project data (Chapter 1) there were over two thousand million models with less than five QTL. Traditional model building algorithms, such as forward selection, have been used with some success for the QTL mapping problem (Chapter 2). However, these only consider a tiny fraction of the potential models and may produce sub-optimal models.

In this thesis a novel selection method, the least absolute selection and shrinkage operator (LASSO; Tibshirani, 1996), is investigated from both the general regression and the QTL mapping contexts. The LASSO effects are incorporated into a mixed model to allow selection in the presence of experimental effects. The extended mixed model appears to be a competitive method for choosing subsets in linear mixed models, especially for the QTL mapping problem.

12.2 Summary of Methodology

A convenient method for circumventing the problem of subset selection is to include all measured genetic information into a single model, an all-marker regression. In QTL mapping experiments the number of genetic markers is often large with respect to the number of observations. Hence, fitting an all-marker model using fixed effects is not sensible. An alternative is to use random effects, or equivalently constrained effects (Whittaker et al., 2000; Gianola et al., 2003; and Xu, 2003). If the random effect all-marker model is to be used then an assumption about the distribution of the random effects has to be made.

Empirical evidence (Chamberlin et al., 2005) suggests that the distribution of QTL effects may have similar kurtosis to that of a double exponential distribution. This implies that assuming that the marker effects follow a double exponential distribution may be a useful model (matching the distribution to the fourth moment). This is the random effects model specification of a novel statistical estimation method called the 'least absolute selection and shrinkage operator' (LASSO; Tibshirani, 1996).

Prior to this thesis the LASSO had not been developed from a random effects model perspective. In particular, the observation that the level of shrinkage can be estimated via estimation of dispersion parameter (Chapter 4) had not been exploited. In this thesis, estimation of dispersion parameters was performed using the marginal likelihood of the observed data. The exact likelihood (Chapter 5) was both intractable and difficult to compute, so two approximate likelihoods were developed (Chapters 7 and 10).

The first approximate likelihood (Chapter 7) formed the basis for the demonstrated QTL mapping methods presented in this thesis. Unfortunately, it produced score equations that were biased. They were adjusted using an empirical estimate of the score equations' means (McCullagh & Tibshirani, 1990). The method for estimating the dispersion parameters for specification of the amount of shrinkage was competitive when compared to other LASSO and subset selection methods.

The second approximation (Chapter 10) was motivated from an alternative specification of the LASSO model. It produced score equations that were approximately unbiased. Hence, they did not require adjusting. The estimates from both the approximate likelihoods agreed well, but the second method required less computation.

Generally, prediction from LASSO models is taken as the mode of the predictive distribution (posterior distribution in Bayesian terminology). Under the random effects model formulation this is not the only choice. In particular, the expected value of the predictive distribution (best predictor) may be superior in terms of mean squared error (Searle et al., 1992). Methods to evaluate these, and other, predictors were given in Chapter 10. A simulation study showed that best predictor had lowest mean squared error but the LASSO predictions were comparable. They also had the benefit of predicting effects to be identically zero.

Two methods of inference were presented in this thesis. The first was based on the hypothesis test $\beta_i = 0$, evaluated via simulation (Chapter 7). This test treated the effects as though they were fixed. In Chapter 10 this was rectified by the presentation of a method to calculate the probabilities $\mathrm{P}(\beta_i > 0|\boldsymbol{y})$ and/or $\mathrm{P}(\beta_i < 0|\boldsymbol{y})$.

So far in this summary, only random effects models containing double exponential effects have been discussed. These models are too simple for analysing QTL mapping data, which require models that allow for experimental effects. This topic was addressed in Chapters 8 and 11 by incorporating LASSO effects into a mixed model containing fixed and random normal effects. It was suggested that dispersion parameters in both models could be estimated using a partial Laplace approximation (Taylor & Verbyla, 2006) to the restricted likelihood. Only the methods described in Chapter 8 were implemented. Simulations showed that the estimation method provided approximately unbiased estimates of dispersion parameters and fixed effects. LASSO, random normal and fixed effects were predicted/estimated using an extension of the mixed model equations (Henderson, 1950).

Incorporating LASSO effects into a mixed model established the methodology to analyse QTL mapping data. The QTL mapping method was defined, assessed via simulation and demonstrated using the Davies' Gene Mapping data in Chapter 9. The new method appeared to be very competitive in locating QTL. For highly heritable traits ($h^2 = 50\%$) it found more QTL without increasing the false positive rate. For lowly heritable traits ($h^2 = 10\%$), a comparable number of QTL were identified but the rate of false positives was lower. The model used in the analysis of the Davies' Gene mapping data included 570 marker-within-sire effects. Only 5 of these were predicted to be non-zero. This showed that the method developed identified areas of the genome that might be associated with the quantitative trait while clearly disregarding the rest.

In summary, the methods presented in this thesis form an approach to LASSO modeling in the presence of experimental effects. For simple models containing only LASSO effects, this approach is at least competitive with other common methods for estimating the amount of shrinkage. The inclusion of fixed and random normal effects into the model generalises its applicability. The extended model (LLMM) is well suited to QTL mapping data due to its distributional assumptions. In simulation studies the LLMM performs better than other commonly used methods and has the ability to predict marker effects to be identically zero.

12.3 Possible Future Research

12.3.1 *Reducing Computation*

A shortcoming of the implemented LLMM model presented in this thesis is the computational effort required to estimate the dispersion parameters. This is mostly due to the stochastic element introduced by computing the bootstrap estimates of the expected scores. If this empirical method could be removed, then computation for the estimation should be greatly reduced.

Implementing the alternative LLMM presented in Chapter 11 should alleviate this problem. However, estimation may still be slow if there are a large number of observations and/or a large number of explanatory variables. Further work will still be needed for a very quick estimation algorithm.

12.3.2 *All-marker Model*

The all-marker model itself may attract some criticism for two reasons. Firstly, the model only considers a finite set of specific positions of the genome rather than the full set of positions. However, if there is a reasonably large number of markers then the loss in power and precision should be minimal compared to an interval mapping approach (e.g. Broman & Speed, 2002). The second potential criticism is that the markers will not be evenly spread throughout the genome. Those markers that are in marker-poor regions *may* or *may not* have a higher probability of being associated with a QTL than those in a marker-rich region. The Bayesian approach taken by Ball (2001) could be used to overcome this. He placed prior information on the markers effects that was based on the density of markers.

12.3.3 *Distributional Assumptions*

The LLMM may not be the most appropriate random effects model for mapping QTL. Other effect distributions could be investigated, for example, the set of distributions used as priors for bridge estimation in Fu (1998). These distributions have a shape parameter as well as a dispersion parameter (the double exponential and normal distributions are members of this family). The shape parameter could be estimated from a marginal likelihood in a similar fashion to the dispersion parameter described in this thesis. However, difficulties are likely to occur when the shape parameter has more mass near zero than the double exponential. In this case, the exponent of the joint distribution of outcomes and random effects will have a non-convex solution space. Hence, calculating the maximum of this function with respect to the random effects will be a non-trivial task.

An alternative method to generalise the distributional assumptions in the LLMM is to generalise the alternative form of the double exponential distribution (Chapter 10). The random variances could be assumed to arise from a more general distribution than the exponential (such as the gamma). If the gamma were chosen, then the shape and dispersion parameters could be estimated from the likelihood of the observed data. Methods to obtain predictions in situations where the shape parameter is less than 1 (the exponential distribution's shape) could again be difficult. Kiiveri (2003) proposed a similar model to this. However, his Bayesian analysis chose a Jeffery's prior in place of the gamma distribution.

Extending the conditional normal representation of the double exponential distribution using a gamma distribution subsumes another class of distributions; namely, the t-distributions with ν degrees of freedom. These can be specified as a conditional normal where the variance arises from a χ^2_ν distribution (e.g. Andrews & Mallows, 1974). The χ^2_ν distribution is a gamma with shape parameter $\nu/2$. The t-distribution is sometimes used in fixed effect models to accommodate large residuals (Lange et al., 1989). This is precisely the nature of marker effects in QTL mapping (although this is an effect distribution and not a residual distribution). However, the t-distribution has a smaller mass of probability near zero than the double exponential and hence may not be as appropriate for QTL mapping.

12.3.4 *Multiple QTL Distributions*

In the analysis of the Davies' Gene Mapping data all the marker-within-sire effects were assumed to come from the same distribution. However, this may not be appropriate. For example, each sire's set of marker effects may exhibit different variance. This can be incorporated by including separate sets of LASSO random effects in the LLMM. Estimation would proceed in a similar manner, but with an extended likelihood. It may also transpire that a single dispersion parameter is not adequate, even within a sire family. For example, the variance of the effects on the different chromosomes may be different. Again, separate dispersion parameters could be fitted, but the model is likely to be difficult to estimate as the number of dispersion effects is large with respect to the number of observations. The individual dispersion parameters could be treated as random, extending the hierarchy of the LLMM. It is not immediately clear what distributional assumption should be made about the dispersion parameters. Perhaps, following the alternative LASSO model, an exponential distribution could be considered. The distribution of the effects, marginal to the proposed random dispersion effects, may be well-approximated by one of the generalisations of the double exponential distribution.

12.3.5 *Multiple Traits*

In most experiments on domestic animals, many traits are measured. Often, the analysis of these experiments considers multiple trait models. If there is genetic or residual correlation between the traits, then these analyses should increase the accuracy of prediction. In QTL mapping experiments, many traits are also recorded. For example, birth weight was just one of approximately 100 traits measured in the Davies' Gene Mapping Project. Multi-trait models are likely to be useful for the analysis of QTL mapping data as they could help identify pleiotropic QTL. There is potential for the LLMM to be extended for the multiple trait analysis. This extension is not available from the standard form of the double exponential distribution, as there appears to be no generalisation that would continue to generate identically zero LASSO predictions. However, the conditional normal representation may hold some hope. The model (ignoring fixed and random normal effects) for the t traits on the i^{th} individual could be

$$\boldsymbol{y}_i = \boldsymbol{B}\boldsymbol{l}_i + \boldsymbol{e}_i$$

where $\boldsymbol{y}_i$, and $\boldsymbol{e}_i$ are the $t \times 1$ column vectors of trait observations and residuals for the i^{th} individual respectively, $\boldsymbol{l}_i$ is the $q \times 1$ column vector of marker observations for the i^{th} individual and $\boldsymbol{B}$ is the $t \times q$ matrix of marker effects. Each of the t rows of $\boldsymbol{B}$ represents the marker effects on a trait. Following the LLMM, it may be appropriate to assume that different marker effects on the same trait are independent. However, the effect of any individual marker may have a correlated effect on different traits. This implies that the elements within a row of $\boldsymbol{B}$ are correlated. Using the conditional normal representation, there are two possible methods to induce such a correlation. One or both may prove to be useful. Let $\boldsymbol{\beta}_k$ be the k^{th} column of $\boldsymbol{B}$ and let $\boldsymbol{\delta}_k$ be the corresponding random variance effects from the conditional normal model.

The first method of inducing a correlation amongst a marker's effects on multiple traits uses the distribution specification

$$\begin{aligned}\boldsymbol{\beta}|\boldsymbol{\delta}_k &\sim \mathrm{N}(\mathbf{0}, \boldsymbol{D}_k^{\frac{1}{2}}\boldsymbol{\Sigma}\boldsymbol{D}_k^{\frac{1}{2}})\\ \delta_{kl} &\sim \exp(2\phi^2), \qquad\qquad \text{for } l = 1\ldots t\end{aligned}$$

where $\boldsymbol{D}_k = \mathrm{diagv}\,(\boldsymbol{\delta}_k)$, $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}(\boldsymbol{\gamma})$ is a $t \times t$ correlation matrix whose structure depends on dispersion parameters $\boldsymbol{\gamma}$. It is expected that the correlation matrix $\boldsymbol{\Sigma}$ should be held constant across all markers.

The second method uses the a conditional normal distribution, but assumes that the $\boldsymbol{\delta}_k$ follow a multivariate distribution. It is not clear what this distribution should be. There are numerous types of multivariate exponentials in Kotz et al. (2000), which may provide a useful representation for the correlation structure. However, a possible method for generalising the exponential to a multivariate distribution may be obtained by first noting that an exponential distribution with mean 2 is the same as a χ^2_2 distribution. A multivariate generalisation of the χ^2_2 distribution is the t-dimensional Wishart distribution with 2 degrees of freedom. In principle, the variances and covariances between the elements of a Wishart distributed $\boldsymbol{\delta}_k$ could be estimated through the expectation specification of the Wishart distribution.

12.3.6 *Outlier Detection*

In this thesis, detection of *important* LASSO effects was performed using inference methods. However, this identification can be considered as outlier detection of the unobserved random effects. There have been many different methods proposed to identify outlying residuals in linear models (e.g. Cook & Weisberg, 1982). Some of these approaches are considered for random effects by Gogel et al. (2001). In particular, the alternative outlier model (AOM) might be useful (Cook et al., 1982; Thompson, 1985; and Verbyla et al., 2006). Using the AOM for LASSO random effects will require special consideration. The distribution assumption for the variance inflation term is not obvious. Consider the simple LASSO model

$$\boldsymbol{y} = \boldsymbol{L}\boldsymbol{\beta} + \boldsymbol{e}$$

where all terms have the same meaning as in Chapter 7. The alternative outlier model for the i^{th} LASSO effect β_i is

$$\boldsymbol{y} = \boldsymbol{L}(\boldsymbol{\beta} + \boldsymbol{d}_i) + \boldsymbol{e}$$

where $\boldsymbol{d}_i$ is a vector of zeros with a non-zero effect ω_i in its i^{th} element. An appropriate assumption to make about the random effects might be that $\{\beta_j\}$ and ω_i are double exponentially distributed variables. However, this would imply that the total effect of the i^{th} explanatory variable $\beta_i + omega_i$ does not have a double exponential distribution. This implies that the AOM is no longer a LASSO model. The AOM with this specification may still be useful for identifying outliers.

12.3.7 *Summary of Future Work*

As presented in this thesis, the LASSO linear mixed model provides a useful model for mapping quantitative trait loci. The topics raised in this discussion are areas for future research. They may broaden and/or strengthen the models applicability and usefulness.

Appendix A

Statistical Results

Result A.1. *(e.g. Flury, 1997) For a jointly normal distributed pair of random variables v_1 and v_2 with distribution*

$$\begin{pmatrix} \boldsymbol{v_1} \\ \boldsymbol{v_2} \end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix} \boldsymbol{\mu_1} \\ \boldsymbol{\mu_2} \end{pmatrix}, \begin{pmatrix} \boldsymbol{V_{11}} & \boldsymbol{V_{12}} \\ \boldsymbol{V_{21}} & \boldsymbol{V_{22}} \end{pmatrix}\right).$$

The conditional expectation and variance of $\boldsymbol{v_1}|\boldsymbol{v_2}$ are

$$\begin{aligned} \mathrm{E}\,(\boldsymbol{v_1}|\boldsymbol{v_2}) &= \boldsymbol{\mu_1} + \boldsymbol{V_{12}}\boldsymbol{V_{22}^{-1}}(\boldsymbol{v_2} - \boldsymbol{\mu_2}) \qquad \textit{and} \\ \mathrm{var}\,(\boldsymbol{v_1}|\boldsymbol{v_2}) &= \boldsymbol{V_{11}} - \boldsymbol{V_{12}}\boldsymbol{V_{22}^{-1}}\boldsymbol{V_{21}}. \end{aligned}$$

Theorem A.1. *The conditional variance matrix of $\boldsymbol{y}_1$ given $\boldsymbol{y}_2$ is $\sigma^2(\boldsymbol{X}^T\boldsymbol{H}^{-1}\boldsymbol{X})^{-1}$.*

Proof. Pre- and post-multiplication of the result in Lemma B.1 by $\boldsymbol{W}_1^T$ and $\boldsymbol{W}_1$ respectively gives

$$\begin{aligned} \boldsymbol{W}_1^T\boldsymbol{H}\boldsymbol{W}_1 - \boldsymbol{W}_1^T\boldsymbol{H}\boldsymbol{W}_2(\boldsymbol{W}_2^T\boldsymbol{H}\boldsymbol{W}_2)^{-1}\boldsymbol{W}_2\boldsymbol{H}\boldsymbol{W}_1 &= \boldsymbol{W}_1^T\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{H}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{W}_1 \\ &= (\boldsymbol{X}^T\boldsymbol{H}^{-1}\boldsymbol{X})^{-1}. \end{aligned}$$

The left hand side is in the form of Result A.1 implying that it is the conditional variance of $\boldsymbol{y}_1$ given $\boldsymbol{y}_2$. This implies that the right hand side is also the same conditional variance. □

Theorem A.2. *In the normal density function of the marginal distribution of $\boldsymbol{y}_2$, the terms involving the transformation $\boldsymbol{W}$ can be expressed free of the transformation by*

1. $|\boldsymbol{W}_2^T\boldsymbol{H}\boldsymbol{W}_2| = |\boldsymbol{W}^T\boldsymbol{W}||\boldsymbol{H}||\boldsymbol{X}^T\boldsymbol{H}^{-1}\boldsymbol{X}|$ *and*

2. $\frac{\boldsymbol{y}_2^T(\boldsymbol{W}_2^T\boldsymbol{H}\boldsymbol{W}_2)^{-1}\boldsymbol{y}_2}{\sigma^2} = \frac{\boldsymbol{y}^T\boldsymbol{P}\boldsymbol{y}}{\sigma^2}$

where $\boldsymbol{P} = \boldsymbol{H}^{-1} - \boldsymbol{H}^{-1}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{H}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{H}^{-1}$.

Proof. 1. Equating the determinants of the variance-covariance matrices specified in the joint distribution to those in the conditional and marginal distributions gives

$$\left|\boldsymbol{W}^T\boldsymbol{H}\boldsymbol{W}\right| = \left|(\boldsymbol{X}^T\boldsymbol{H}^{-1}\boldsymbol{X})^{-1}\right|\left|\boldsymbol{W}_2^T\boldsymbol{H}\boldsymbol{W}_2\right|.$$

Rearranging gives

$$\left|\boldsymbol{W}_2^T\boldsymbol{H}\boldsymbol{W}_2\right| = \left|\boldsymbol{W}\boldsymbol{W}^T\right|\left|\boldsymbol{H}\right|\left|\boldsymbol{X}\boldsymbol{H}^{-1}\boldsymbol{X}\right|.$$

2. Consider a rearrangement of the result in Lemma B.1, namely

$$\boldsymbol{H}\boldsymbol{W}_2(\boldsymbol{W}_2^T\boldsymbol{H}\boldsymbol{W}_2)^{-1}\boldsymbol{W}_2^T\boldsymbol{H} = \boldsymbol{H} - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{H}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T.$$

Pre- and post-multiply both sides by $\boldsymbol{H}^{-1}$ gives

$$\begin{aligned}\boldsymbol{W}_2(\boldsymbol{W}_2^T\boldsymbol{H}\boldsymbol{W}_2)^{-1}\boldsymbol{W}_2^T &= \boldsymbol{H}^{-1} - \boldsymbol{H}^{-1}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{H}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{H}^{-1}.\\ &= \boldsymbol{P} \qquad\qquad \text{say.}\end{aligned}$$

Writing $\boldsymbol{y}_2$ as $\boldsymbol{W}_2^T\boldsymbol{y}$ gives an expression for the numerator in the exponent

$$\begin{aligned}\boldsymbol{y}_2^T(\boldsymbol{W}_2^T\boldsymbol{H}\boldsymbol{W}_2)^{-1}\boldsymbol{y}_2/\sigma^2 &= \boldsymbol{y}^T\boldsymbol{W}_2(\boldsymbol{W}_2^T\boldsymbol{H}\boldsymbol{W}_2)^{-1}\boldsymbol{W}_2\boldsymbol{y}/\sigma^2\\ &= \boldsymbol{y}^T\boldsymbol{P}\boldsymbol{y}/\sigma^2.\end{aligned}$$

□

Theorem A.3. *(Cullis et al., 2006) The score equation for η_i from the restricted likelihood in Chapter 3 is*

$$U_p(\eta_i) = -\frac{1}{2}\left(tr\left(\boldsymbol{P}\dot{\boldsymbol{H}}_i\right) - \boldsymbol{y}^T\boldsymbol{P}\dot{\boldsymbol{H}}_i\boldsymbol{P}\boldsymbol{y}/\sigma^2\right).$$

Proof. First consider the derivative of the log-determinants and use Result B.1

$$\begin{aligned}\frac{\partial\log|\boldsymbol{H}|}{\partial\eta_i} + \frac{\partial\log|\boldsymbol{X}^T\boldsymbol{H}^{-1}\boldsymbol{X}|}{\partial\eta_i} &= \text{tr}\left(\boldsymbol{H}^{-1}\dot{\boldsymbol{H}}_{\boldsymbol{i}}\right) + \text{tr}\left((\boldsymbol{X}^T\boldsymbol{H}^{-1}\boldsymbol{X})^{-1}\frac{\partial\boldsymbol{X}^T\boldsymbol{H}^{-1}\boldsymbol{X}}{\partial\eta_i}\right)\\ &= \text{tr}\left(\boldsymbol{H}^{-1}\dot{\boldsymbol{H}}_{\boldsymbol{i}}\right) - \text{tr}\left((\boldsymbol{X}^T\boldsymbol{H}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{H}^{-1}\dot{\boldsymbol{H}}_{\boldsymbol{i}}\boldsymbol{H}^{-1}\boldsymbol{X}\right)\\ &= \text{tr}\left(\boldsymbol{H}^{-1}\dot{\boldsymbol{H}}_{\boldsymbol{i}} - (\boldsymbol{X}^T\boldsymbol{H}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{H}^{-1}\dot{\boldsymbol{H}}_{\boldsymbol{i}}\boldsymbol{H}^{-1}\boldsymbol{X}\right)\\ &= \text{tr}\left(\boldsymbol{H}^{-1}\dot{\boldsymbol{H}}_{\boldsymbol{i}} - \boldsymbol{H}^{-1}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{H}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{H}^{-1}\dot{\boldsymbol{H}}_{\boldsymbol{i}}\right)\\ &= \text{tr}\left(\boldsymbol{P}\dot{\boldsymbol{H}}_{\boldsymbol{i}}\right).\end{aligned}$$

Now consider the sums of squares in the restricted log-likelihood and use Result B.1

$$\begin{aligned}\frac{\partial\boldsymbol{y}^T\boldsymbol{P}\boldsymbol{y}}{\partial\eta_i} &= \boldsymbol{y}^T\frac{\partial\boldsymbol{P}}{\partial\eta_i}\boldsymbol{y}\\ &= -\boldsymbol{y}^T\boldsymbol{P}\dot{\boldsymbol{H}}\boldsymbol{P}\boldsymbol{y}.\end{aligned}$$

□

Theorem A.4. *(Seber, 1977) Let $\boldsymbol{z}$ be an $n \times 1$ vector of random variables and let $\boldsymbol{A}$ be an $n \times n$ symmetric matrix. If* $\mathrm{E}\left(\boldsymbol{z}\right) = \boldsymbol{\theta}$ *and* $\mathrm{var}\left(\boldsymbol{z}\right) = \boldsymbol{\Sigma}$*, then*

$$\mathrm{E}\left(\boldsymbol{z}^T\boldsymbol{A}\boldsymbol{z}\right) = tr[\boldsymbol{A}\boldsymbol{\Sigma}] + \boldsymbol{\theta}^T\boldsymbol{A}\boldsymbol{\theta}.$$

Proof.

$$\begin{aligned}
\mathrm{E}\left(\boldsymbol{z}^T\boldsymbol{A}\boldsymbol{z}\right) &= \mathrm{E}\left((\boldsymbol{z}-\boldsymbol{\theta})^T\boldsymbol{A}(\boldsymbol{z}-\boldsymbol{\theta}) + 2\boldsymbol{z}^T\boldsymbol{A}\boldsymbol{\theta} - \boldsymbol{\theta}^T\boldsymbol{A}\boldsymbol{\theta}\right) \\
&= \mathrm{E}\left((\boldsymbol{z}-\boldsymbol{\theta})^T\boldsymbol{A}(\boldsymbol{z}-\boldsymbol{\theta})\right) + 2\mathrm{E}\left(\boldsymbol{z}^T\boldsymbol{A}\boldsymbol{\theta}\right) - \boldsymbol{\theta}^T\boldsymbol{A}\boldsymbol{\theta} \\
&= \sum_i\sum_j a_{ij}\mathrm{E}\left((z_i-\theta_i)(z_j-\theta_j)\right) + \boldsymbol{\theta}^T\boldsymbol{A}\boldsymbol{\theta} \\
&= \sum_i\sum_j a_{ij}\sigma_{ij} + \boldsymbol{\theta}^T\boldsymbol{A}\boldsymbol{\theta} \\
&= tr[\boldsymbol{A}\boldsymbol{\Sigma}] + \boldsymbol{\theta}^T\boldsymbol{A}\boldsymbol{\theta}.
\end{aligned}$$

□

Theorem A.5. *The elements of the observed information matrix from the restricted log-likelihood are*

$$\begin{aligned}
\mathcal{I}_o(\sigma^2,\sigma^2) &= -\frac{(n-p)}{2\sigma^4} + \frac{\boldsymbol{y}^T\boldsymbol{P}\boldsymbol{y}}{\sigma^6} \\
\mathcal{I}_o(\sigma^2,\eta_i) &= \frac{\boldsymbol{y}^T\boldsymbol{P}\dot{\boldsymbol{H}}_i\boldsymbol{P}\boldsymbol{y}}{2\sigma^4} \\
\mathcal{I}_o(\eta_i,\eta_j) &= \frac{1}{2}\mathrm{tr}\left(\boldsymbol{P}\ddot{\boldsymbol{H}}_{ij}\right) - \frac{1}{2}\mathrm{tr}\left(\boldsymbol{P}\dot{\boldsymbol{H}}_i\boldsymbol{P}\dot{\boldsymbol{H}}\right) + \frac{\boldsymbol{y}^T\boldsymbol{P}\dot{\boldsymbol{H}}_i\boldsymbol{P}\dot{\boldsymbol{H}}\boldsymbol{P}\boldsymbol{y}}{\sigma^2} - \frac{\boldsymbol{y}^T\boldsymbol{P}\ddot{\boldsymbol{H}}_{ij}\boldsymbol{P}\boldsymbol{y}}{2\sigma^2}
\end{aligned}$$

where $\ddot{\boldsymbol{H}}_{ij} = \frac{\partial^2\boldsymbol{H}}{\partial\eta_i\partial\eta_j}$.

Proof. The observed information matrix is the negative of the Hessian for the restricted likelihood function. Using the matrix derivatives in Result B.1 the Hessian has elements

$$\begin{aligned}
\frac{\partial^2\ell_r}{\partial\sigma^4} &= \frac{\partial U(\sigma^2)}{\partial\sigma^2} \\
&= \frac{(n-p)}{2\sigma^4} - \frac{\boldsymbol{y}^T\boldsymbol{P}\boldsymbol{y}}{\sigma^6}, \\
\frac{\partial^2\ell_r}{\partial\sigma^2\partial\eta_i} &= \frac{\partial U(\sigma^2)}{\partial\eta_i} \\
&= -\frac{\boldsymbol{y}^T\boldsymbol{P}\dot{\boldsymbol{H}}_i\boldsymbol{P}\boldsymbol{y}}{2\sigma^4}
\end{aligned}$$

and

$$
\begin{aligned}
\frac{\partial^2 \ell_r}{\partial \eta_i \partial \eta_j} &= \frac{\partial U(\eta_i)}{\partial \eta_j} \\
&= -\frac{1}{2}\left(\frac{\partial \mathrm{tr}\left(\boldsymbol{P}\dot{\boldsymbol{H}}_i\right)}{\partial \eta_j} - \frac{1}{\sigma^2}\frac{\partial \boldsymbol{y}^T\boldsymbol{P}\dot{\boldsymbol{H}}_i\boldsymbol{P}\boldsymbol{y}}{\partial \eta_j}\right) \\
&= -\frac{1}{2}\left(\mathrm{tr}\left(-\boldsymbol{P}\dot{\boldsymbol{H}}_j\boldsymbol{P}\dot{\boldsymbol{H}}_i + \boldsymbol{P}\ddot{\boldsymbol{H}}_{ij}\right) + \boldsymbol{y}^T\left(\boldsymbol{P}\ddot{\boldsymbol{H}}_{ij}\boldsymbol{P} - 2\boldsymbol{P}\dot{\boldsymbol{H}}_i\boldsymbol{P}\dot{\boldsymbol{H}}_j\boldsymbol{P}\right)\boldsymbol{y}/\sigma^2\right) \\
&= -\frac{1}{2}\mathrm{tr}\left(\boldsymbol{P}\ddot{\boldsymbol{H}}_{ij}\right) + \frac{1}{2}\mathrm{tr}\left(\boldsymbol{P}\dot{\boldsymbol{H}}_j\boldsymbol{P}\dot{\boldsymbol{H}}_i\right) - \frac{1}{\sigma^2}\boldsymbol{y}^T\boldsymbol{P}\dot{\boldsymbol{H}}_i\boldsymbol{P}\dot{\boldsymbol{H}}_j\boldsymbol{P}\boldsymbol{y} + \frac{1}{2\sigma^2}\boldsymbol{y}^T\boldsymbol{P}\ddot{\boldsymbol{H}}_{ij}\boldsymbol{P}\boldsymbol{y}.
\end{aligned}
$$

□

Theorem A.6. *The elements of the expected information matrix from the restricted log-likelihood are*

$$
\begin{aligned}
\mathcal{I}_e(\sigma^2, \sigma^2) &= \frac{(n-p)}{2\sigma^4} \\
\mathcal{I}_e(\sigma^2, \eta_i) &= \frac{1}{2\sigma^2}\mathrm{tr}\left(\boldsymbol{P}\dot{\boldsymbol{H}}_i\right) \\
\mathcal{I}_e(\eta_i, \eta_j) &= \frac{1}{2}\mathrm{tr}\left(\boldsymbol{P}\dot{\boldsymbol{H}}_i\boldsymbol{P}\dot{\boldsymbol{H}}_j\right).
\end{aligned}
$$

Proof. Note that $\boldsymbol{PX} = \boldsymbol{0}$ and $\boldsymbol{PHP} = \boldsymbol{P}$. Using the results for expected values of quadratic forms in Result A.4 the expected values of the quadratic forms in the observed information in Theorem A.5 are

$$
\begin{aligned}
\mathrm{E}\left(\boldsymbol{y}^T\boldsymbol{P}\boldsymbol{y}\right) &= \sigma^2\mathrm{tr}\left(\boldsymbol{PH}\right) + \boldsymbol{\tau}^T\boldsymbol{X}^T\boldsymbol{P}\boldsymbol{X}\boldsymbol{\tau} \\
&= \sigma^2\mathrm{tr}\left(\boldsymbol{I}_n\right) - \sigma^2\mathrm{tr}\left(\boldsymbol{H}^{-1}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{H}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\right) \\
&= \sigma^2(n-p) \qquad\qquad \text{as second term is rank } p, \\
\mathrm{E}\left(\boldsymbol{y}^T\boldsymbol{P}\dot{\boldsymbol{H}}_i\boldsymbol{P}\boldsymbol{y}\right) &= \sigma^2\mathrm{tr}\left(\boldsymbol{P}\dot{\boldsymbol{H}}_i\boldsymbol{P}\boldsymbol{H}\right) + \boldsymbol{\tau}^T\boldsymbol{X}^T\boldsymbol{P}\dot{\boldsymbol{H}}_i\boldsymbol{P}\boldsymbol{X}\boldsymbol{\tau} \\
&= \sigma^2\mathrm{tr}\left(\boldsymbol{P}\dot{\boldsymbol{H}}_i\right), \\
\mathrm{E}\left(\boldsymbol{y}^T\boldsymbol{P}\dot{\boldsymbol{H}}_i\boldsymbol{P}\dot{\boldsymbol{H}}_j\boldsymbol{P}\boldsymbol{y}\right) &= \sigma^2\mathrm{tr}\left(\boldsymbol{P}\dot{\boldsymbol{H}}_i\boldsymbol{P}\dot{\boldsymbol{H}}_j\boldsymbol{P}\boldsymbol{H}\right) + \boldsymbol{\tau}^T\boldsymbol{X}^T\boldsymbol{P}\dot{\boldsymbol{H}}_i\boldsymbol{P}\dot{\boldsymbol{H}}_j\boldsymbol{P}\boldsymbol{X}\boldsymbol{\tau} \\
&= \sigma^2\mathrm{tr}\left(\boldsymbol{P}\dot{\boldsymbol{H}}_i\boldsymbol{P}\dot{\boldsymbol{H}}_j\right) \qquad\qquad \text{and} \\
\mathrm{E}\left(\boldsymbol{y}^T\boldsymbol{P}\ddot{\boldsymbol{H}}_i\boldsymbol{P}\boldsymbol{y}\right) &= \sigma^2\mathrm{tr}\left(\boldsymbol{P}\ddot{\boldsymbol{H}}_i\boldsymbol{P}\boldsymbol{H}\right) + \boldsymbol{\tau}^T\boldsymbol{X}^T\boldsymbol{P}\ddot{\boldsymbol{H}}_i\boldsymbol{P}\boldsymbol{X}\boldsymbol{\tau} \\
&= \sigma^2\mathrm{tr}\left(\boldsymbol{P}\ddot{\boldsymbol{H}}_i\right).
\end{aligned}
$$

These expectations are then substituted into the observed information elements in Theorem A.5 to produce the result. □

Theorem A.7. *The moment generating function of the central Laplace distribution with dispersion parameter ϕ is*

$$
m(t) = \left(1 - t^2\phi^2\right)^{-1}.
$$

Proof. The functional form of the Laplace distribution is

$$f(y) = \frac{1}{2\phi}\exp\left(-\frac{1}{\phi}|y|\right)$$

The moment generating function is

$$\begin{aligned}
m(t) &= \mathrm{E}\left(\exp\left(ty\right)\right) \\
&= \int_{-\infty}^{\infty} \exp\left(ty\right)\frac{1}{2\phi}\exp\left(-\frac{1}{\phi}|y|\right)dy \\
&= \frac{1}{2\phi}\left(\int_{0}^{\infty}\exp\left(-\frac{y}{\phi/(1-t\phi)}\right)dy + \int_{-\infty}^{0}\exp\left(\frac{y}{\phi/(1+t\phi)}\right)\right) \\
&= \frac{1}{2\phi}\left(\frac{\phi}{1-t\phi}+\frac{\phi}{1+t\phi}\right) \\
&= \left(1-t^2\phi^2\right)^{-1}.
\end{aligned}$$

□

Corollary A.1. *The variance of a central Laplace distribution with dispersion parameter ϕ is $2\phi^2$.*

Proof. The first two derivatives of the moment generating function are

$$\begin{aligned}
\frac{dm(t)}{dt} &= (1-t^2\phi^2)^{-2}(2t\phi^2) \\
&= \frac{2t\phi^2}{(1-t^2\phi^2)^2} \\
\frac{d^2m(t)}{dt^2} &= \frac{8t^2\phi^4}{(1-t^2\phi^2)^3}+\frac{2\phi^2}{(1-t^2\phi^2)}.
\end{aligned}$$

Evaluating the second derivative at $t=0$ gives the variance

$$\mathrm{var}\left(y\right) = \frac{0}{1}+\frac{2\phi^2}{1} = 2\phi^2.$$

□

Definition A.1. *(e.g. Mardia et al., 1979) The singular multivariate normal density function is given by*

$$f(\boldsymbol{y}) = \frac{(2\pi)^{-\kappa/2}}{(\lambda_1\cdots\lambda_\kappa)^{1/2}}\exp\left(-\frac{1}{2}\left(\boldsymbol{y}-\boldsymbol{\mu}\right)^T\boldsymbol{\Sigma}^-\left(\boldsymbol{y}-\boldsymbol{\mu}\right)\right)$$

where $\boldsymbol{\Sigma}^-$ is a generalised inverse of the matrix $\boldsymbol{\Sigma}$ of rank κ, $(\lambda_1,\ldots,\lambda_\kappa)$ are the non-zero eigenvalues of $\boldsymbol{\Sigma}$ and $\boldsymbol{\mu}$ is the mean vector.

Appendix B

Vector and Matrix Algebra Results

Lemma B.1. *(Verbyla, 1990) If the matrices $\boldsymbol{X}$ and $\boldsymbol{W}_2$ are such that $\boldsymbol{W}_2^T\boldsymbol{X} = \boldsymbol{0}$ and $\boldsymbol{H}$ is positive definite then*

$$\boldsymbol{H} - \boldsymbol{H}\boldsymbol{W}_2(\boldsymbol{L}_2^T\boldsymbol{H}\boldsymbol{L}_2)^{-1}\boldsymbol{W}_2\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{H}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T.$$

Proof. Transform the standard linear mixed model by $\boldsymbol{H}^{-\frac{1}{2}}$; the inverse symmetric square root of $\boldsymbol{H}$. Define $\boldsymbol{X}^* = \boldsymbol{H}^{-\frac{1}{2}}\boldsymbol{X}$. Then, a basis for the orthogonal complement of $\mathcal{R}(\boldsymbol{X}^*)$ is given by the columns of $\boldsymbol{W}_2^* = \boldsymbol{H}^{\frac{1}{2}}\boldsymbol{W}_2$ as

$$\begin{aligned}\boldsymbol{W}_2^{*T}\boldsymbol{X}^* &= \boldsymbol{W}_2^T\boldsymbol{H}^{\frac{1}{2}}\boldsymbol{H}^{-\frac{1}{2}}\boldsymbol{X}\\ &= \boldsymbol{W}_2^T\boldsymbol{X}\\ &= \boldsymbol{0}.\end{aligned}$$

Since $\mathcal{R}(\boldsymbol{W}_2^*)$ is orthogonal to $\mathcal{R}(\boldsymbol{X}^*)$ then it is possible to express the orthogonal projection onto $\mathcal{R}(\boldsymbol{W}_2^*)$ by (Seber, 1977)

$$\boldsymbol{W}_2^*(\boldsymbol{W}_2^{*T}\boldsymbol{W}_2^*)^{-1}\boldsymbol{W}_2^{*T} = \boldsymbol{I_n} - \boldsymbol{X}^*(\boldsymbol{X}^{*T}\boldsymbol{X}^*)^{-1}\boldsymbol{X}^{*T}$$

which is equivalent to

$$\boldsymbol{H}^{\frac{1}{2}}\boldsymbol{W}_2(\boldsymbol{W}_2^T\boldsymbol{H}\boldsymbol{W}_2)^{-1}\boldsymbol{W}_2\boldsymbol{H}^{\frac{1}{2}} = \boldsymbol{I_n} - \boldsymbol{H}^{-\frac{1}{2}}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{H}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{H}^{-\frac{1}{2}}.$$

Multiplication of both sides by $\boldsymbol{H}^{\frac{1}{2}}$ yields the equation in the Lemma statement. □

Result B.1. *(e.g. Cullis et al., 2006) Let $\boldsymbol{A}$ be a $n \times n$ symmetric matrix and $\boldsymbol{B}$ a $p \times p$ matrix that is: a function of $\boldsymbol{\eta}$ and is non-singular. With $\dot{\boldsymbol{B}}_i = \frac{\partial \boldsymbol{B}}{\partial \eta_i}$ the following hold*

1. $\frac{\partial \boldsymbol{a}^T\boldsymbol{x}}{\partial \boldsymbol{x}} = \boldsymbol{a}$

2. $\frac{\partial \boldsymbol{x}^T\boldsymbol{x}}{\partial \boldsymbol{x}} = 2\boldsymbol{x}, \qquad \frac{\partial \boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x}}{\partial \boldsymbol{x}} = 2\boldsymbol{A}\boldsymbol{x}, \qquad \frac{\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{y}}{\partial \boldsymbol{x}} = \boldsymbol{A}\boldsymbol{y}$

3. $\frac{\partial \log|\boldsymbol{B}|}{\partial \eta_i} = \text{tr}\left(\boldsymbol{B^{-1}}\dot{\boldsymbol{B}}_{\boldsymbol{i}}\right)$

4. $\frac{\partial \boldsymbol{B^{-1}}}{\partial \eta_i} = -\boldsymbol{B^{-1}}\dot{\boldsymbol{B}}_{\boldsymbol{i}}\boldsymbol{B^{-1}}$

5. $\frac{\partial \boldsymbol{P}}{\partial \kappa_i} = -\boldsymbol{P}\dot{\boldsymbol{H}}_i\boldsymbol{P}$, *where* $\boldsymbol{P} = \boldsymbol{H}^{-1} - \boldsymbol{H}^{-1}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{H}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{H}^{-1}$.

Result B.2. *(Cullis et al., 2006) For non-singular matrices* $\boldsymbol{R}$, $\boldsymbol{G}$ *and* $\boldsymbol{R}$ *of dimensions* $n \times n$, $q \times q$ *and for* $n \times q$ *respectively.*

$$(\boldsymbol{R} + \boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}^T)^{-1} = \boldsymbol{R}^{-1} - \boldsymbol{R}^{-1}\boldsymbol{Z}(\boldsymbol{G}^{-1} + \boldsymbol{Z}^T\boldsymbol{R}^{-1}\boldsymbol{Z})^{-1}\boldsymbol{Z}^T\boldsymbol{R}^{-1}$$

if and only if

$$(\boldsymbol{G}^{-1} + \boldsymbol{Z}^T\boldsymbol{R}^{-1}\boldsymbol{Z})^{-1}\boldsymbol{Z}^T\boldsymbol{R}^{-1} = \boldsymbol{G}\boldsymbol{Z}^T(\boldsymbol{R} + \boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}^T)^{-1}.$$

This follows from the identity

$$\begin{aligned}\boldsymbol{H}^{-1} &= (\boldsymbol{Z}\boldsymbol{G}\boldsymbol{Z}^T + \boldsymbol{R})^{-1}\\ &= \boldsymbol{R}^{-1} - \boldsymbol{R}^{-1}\boldsymbol{Z}(\boldsymbol{Z}^T\boldsymbol{R}^{-1}\boldsymbol{Z} + \boldsymbol{G}^{-1})^{-1}\boldsymbol{Z}^T\boldsymbol{R}^{-1}.\end{aligned}$$

Result B.3. *For the* $p \times n$ *matrix* $\boldsymbol{D}$, *the* $n \times p$ *matrix* $\boldsymbol{E}$ *and the non-singular* $p \times p$ *matrix* $\boldsymbol{A}$

$$\begin{aligned}|\boldsymbol{A} + \boldsymbol{D}\boldsymbol{E}| &= |\boldsymbol{A}||\boldsymbol{I}_p + \boldsymbol{A}^{-1}\boldsymbol{D}\boldsymbol{E}|\\ &= |\boldsymbol{A}||\boldsymbol{I}_n + \boldsymbol{E}\boldsymbol{A}^{-1}\boldsymbol{D}|.\end{aligned}$$

Result B.4. *(Harville, 1997) The inverse of a partitioned matrix* $\boldsymbol{A} = \begin{pmatrix}\boldsymbol{A}_{11} & \boldsymbol{A}_{12}\\ \boldsymbol{A}_{21} & \boldsymbol{A}_{22}\end{pmatrix}$, *assuming that all necessary inverses exist, is*

$$\boldsymbol{A}^{-1} = \begin{pmatrix}\boldsymbol{A}^{11} & \boldsymbol{A}^{12}\\ \boldsymbol{A}^{21} & \boldsymbol{A}^{22}\end{pmatrix} = \begin{pmatrix}\boldsymbol{T} & -\boldsymbol{T}\boldsymbol{A}_{12}\boldsymbol{A}_{22}^{-1}\\ -\boldsymbol{A}_{22}^{-1}\boldsymbol{A}_{21}\boldsymbol{T} & \boldsymbol{A}_{22}^{-1} + \boldsymbol{A}_{22}^{-1}\boldsymbol{A}_{21}\boldsymbol{T}\boldsymbol{A}_{12}\boldsymbol{A}_{22}^{-1}\end{pmatrix}$$

where $\boldsymbol{T} = (\boldsymbol{A}_{11} - \boldsymbol{A}_{12}\boldsymbol{A}_{22}^{-1}\boldsymbol{A}_{21})^{-1}$.

Theorem B.1. *(Harville, 1997, Theorem 9.6.1) Let* $\boldsymbol{T}$ *represent an* $m \times p$ *matrix,* $\boldsymbol{U}$ *an* $m \times q$ *matrix,* $\boldsymbol{V}$ *an* $n \times p$ *matrix, and* $\boldsymbol{W}$ *an* $n \times q$ *matrix, and define* $\boldsymbol{Q} = \boldsymbol{W} - \boldsymbol{V}\boldsymbol{T}^{-}\boldsymbol{U}$. *Suppose that the column space of* $\boldsymbol{U}$ *is a subspace of the column space of* $\boldsymbol{T}$ *and that the row space of* $\boldsymbol{V}$ *is a subspace of the row space of* $\boldsymbol{T}$. *Then, the partitioned matrix*

$$\begin{pmatrix}\boldsymbol{T}^{-} + \boldsymbol{T}^{-}\boldsymbol{U}\boldsymbol{Q}^{-}\boldsymbol{V}\boldsymbol{T}^{-} & -\boldsymbol{T}^{-}\boldsymbol{U}\boldsymbol{Q}^{-}\\ -\boldsymbol{Q}^{-}\boldsymbol{V}\boldsymbol{T}^{-} & \boldsymbol{Q}^{-}\end{pmatrix} = \begin{pmatrix}\boldsymbol{T}^{-} & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{0}\end{pmatrix} + \begin{pmatrix}-\boldsymbol{T}^{-}\boldsymbol{U}\\ \boldsymbol{I}_q\end{pmatrix}\boldsymbol{Q}^{-}\left(-\boldsymbol{V}\boldsymbol{T}^{-}, \boldsymbol{I}_q\right)$$

is a generalised inverse of $\begin{pmatrix}\boldsymbol{T} & \boldsymbol{U}\\ \boldsymbol{V} & \boldsymbol{W}\end{pmatrix}$.

Theorem B.2. *(Harville, 1997, Theorem 9.6.5) Let $\boldsymbol{T}$, $\boldsymbol{U}$, $\boldsymbol{V}$, $\boldsymbol{W}$ and $\boldsymbol{Q}$ be as defined in Theorem B.1 with the same row and column space properties. Then, for any generalised inverse $\boldsymbol{B} = \begin{pmatrix} \boldsymbol{B}_{11} & \boldsymbol{B}_{12} \\ \boldsymbol{B}_{21} & \boldsymbol{B}_{22} \end{pmatrix}$ of the partitioned matrix $\begin{pmatrix} \boldsymbol{T} & \boldsymbol{U} \\ \boldsymbol{V} & \boldsymbol{W} \end{pmatrix}$, the $q \times n$ submatrix $\boldsymbol{B}_{22}$ is a generalised inverse of $\boldsymbol{Q}$.*

Appendix C

Miscellaneous Results

Definition C.1. *For a n dimensional vector $\boldsymbol{x}$ the L_p-norm for $0 \leq p < \infty$ and the L_∞-norm are defined as*

$$\|\boldsymbol{x}\|_p = \sum_{i=1}^{n} |x_i|^p \qquad \textit{and}$$

$$\|\boldsymbol{x}\|_\infty = \underset{0 \leq i \leq n}{max} \; |x_i|$$

The L_1-norm and the L_2-norm are special cases of the L_p-norm for $p = 1$ and $p = 2$ respectively.

Definition C.2. *(Osborne, 1985) Define the function $f : \mathcal{X} \rightarrow \mathcal{Y}$ where $\mathcal{X} \subseteq \mathbb{R}^k$ and $\mathcal{Y} \subseteq \mathbb{R}$. The k dimensional vector $\boldsymbol{d}$ is a sub-gradient of f at $\boldsymbol{x} \in \mathcal{X}$ if and only if for all $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{X}$*

$$f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \boldsymbol{d}^T(\boldsymbol{y} - \boldsymbol{x}).$$

Geometrically, a sub-gradient d at $\boldsymbol{x}$ is a tangent of f at $\boldsymbol{x}$ such that the line through $f(\boldsymbol{x})$ with gradient d never enters the graph of f.

Definition C.3. *(Osborne, 1985) For the function f defined in Definition C.2 the set of all sub-gradients at a point $\boldsymbol{x} \in \mathcal{X}$ is called the sub-differential of f at $\boldsymbol{x}$ and has notation $\partial(\boldsymbol{x})$. When f is multivariate the partial sub-differential is given by $\partial_{\boldsymbol{x}_1}(\boldsymbol{x}_1, \boldsymbol{x}_2)$.*

Result C.1. *(Nash & Sofer, 1996) Consider a non-linear constraint problem with objective function $s(\boldsymbol{x})$ and constraints $\boldsymbol{t}(\boldsymbol{x})$. Define $\boldsymbol{N}(\boldsymbol{x})$ such that its columns form a basis for the null space of the Jacobian of the constraints at $\boldsymbol{x}$ and $\mathscr{L}(\boldsymbol{x}, \boldsymbol{\lambda})$ to be the Lagrangian. The constraint problem has an optimal solution at $\boldsymbol{x}_*$ with associated Lagrange multipliers $\boldsymbol{\lambda}_*$ if the Karush-Kuhn-Tucker conditions are met. These are*

- $\frac{\partial \mathscr{L}(\boldsymbol{x}_*, \boldsymbol{\lambda}_*)}{\partial \boldsymbol{x}} = \boldsymbol{0}$
- $\boldsymbol{\lambda}_* \geq \boldsymbol{0}$

- $\boldsymbol{\lambda}_*^T t(\boldsymbol{x}_*) = \mathbf{0}$
- $\boldsymbol{N}(\boldsymbol{x}_*)^T \frac{\partial^2 \mathscr{L}(\boldsymbol{x},\boldsymbol{\lambda})}{\partial \boldsymbol{x} \partial \boldsymbol{x}^T} \boldsymbol{N}(\boldsymbol{x}_*)$ *is positive semi-definite.*

The first point reflects the first-order optimality conditions, the second is the non-negative constraint on the Lagrange multipliers, the third imposes that if the constraint function is inactive then the associated Lagrange multiplier is zero and the fourth condition stipulates that the stationary point must be a minimum.

Appendix D

S-PLUS Functions

These are S functions to perform the methods in the main body of the thesis. Use them with love and caution. Please advise me if you find any mistakes.

4.1 Fitting the Constrained Regression for Correlated Observations

See Chapter 6 for details.

```
#------------------------------------------------------------------ Constrained Prediction
lasso.t<-function(LPL,LPy,t=1,fm.prev=NULL,print.stuff=F)
#-----------------------------------------------------------------------------------------
#Calculates LASSO constrained regression predictions.

#L is the LASSO design matrix, P is variance matrix of outcomes LPL=t(L)%*%P%*%L, y is the
#vector of outcomes, LPy=t(L)%*%P%*%y, t is the constraint, fm.prev is a previous fit of
#the LASSO constraint model (with smaller constraint) and print.stuff provides an iteration
#trace for the function.

#function returns a list whose elements are the predictions, the corresponding v, lambda
#(eigen value) for the given t, various quantities for the lasso model (for later fitting)
#and the set of indexes for the non-zero predictions.
#-----------------------------------------------------------------------------------------
{

  #---------------------------------------------------------- Initialises common quantities
  init.const<-function(LPL,LPy,fm.prev=NULL){
    if (is.null(fm.prev))
      list(LRL.inv=NULL,LRL=NULL,LRy=NULL,LRL.full=LPL,LRy.full=LPy)
    else
      fm.prev$const
  }

  #------------------------------------------------------------------------- v Calculation
  calc.v<-function(const,b,set){
    if (length(set)==0)
      calc.v<-const$LRy.full
    else{
      b.s<-as.matrix(b[set])
      lambda.inf<-max(abs(const$LRy-const$LRL%*%b.s))
      if (lambda.inf==0)
        (const$LRy.full-const$LRL.full%*%as.matrix(b))
      else
      (const$LRy.full-const$LRL.full%*%as.matrix(b))/lambda.inf
    }
  }

  #------------------------------------------------- Checks if current solution is optimal
  is.optimal<-function(v.vec,set){
```

```
  flag<-T
  if (length(set)==0)
    flag<-F
  else
    if (length(v.vec)>length(set))
      if (max(abs(v.vec[setdiff(1:length(v.vec),set)]))>1)
        flag<-F
  flag
}

#--------------------------------------------------- Finds next regressor to enter model
get.regressor<-function(vec,set){
  not.active<-setdiff(1:length(vec),set)
  pos<-order(-abs(vec[not.active]))[1]
  c(set,not.active[pos])
}

#--------------------------------------- Calculates the sign of the non-zero predictions
get.sign<-function(b,set,vec){
  b.s<-b[set]
  temp<-sign(b.s)
  for (ii in 1:length(set))
    if (temp[ii]==0)
      temp[ii]<- sign(vec[set[ii]])
  temp
}

#----------------------------------------Updates common quantities for current iteration
update.const<-function(old.const,new.S,old.S=c()){
  if (!(length(old.S)<1)){
    if (length(new.S)>length(old.S)){
      #LRL^-1
      loc<-setdiff(new.S,old.S)
      LRL.inv<-old.const$LRL.inv
      LRz<-old.const$LRL.full[old.S,loc]
      zRz<-old.const$LRL.full[loc,loc]
      zPz<-zRz-t(LRz)%*%LRL.inv%*%LRz
      co<--LRL.inv%*%LRz%*%(1/zPz)
      new.mat<-rbind(LRL.inv+LRL.inv%*%LRz%*%(1/zPz)%*%t(LRz)%*%LRL.inv,t(co))
      new.mat<-cbind(new.mat,c(co,(1/zPz)))

      new.LRL<-old.const$LRL.full[new.S,new.S]
      new.LRy<-old.const$LRy.full[new.S]
    }
    else{
      #LRL & LRy
      new.LRL<-old.const$LRL.full[new.S,new.S]
      new.mat<-solve(new.LRL)
      new.LRy<-old.const$LRy.full[new.S]
      dim<-length(new.mat[1,])
    }
  }
  else{
    new.LRL<-old.const$LRL.full[new.S,new.S]
    new.LRy<-old.const$LRy.full[new.S]
    new.mat<-1/new.LRL
  }
list(LRL.inv=new.mat,LRL=new.LRL,LRy=new.LRy,LRL.full=old.const$LRL.full,LRy.full=old.const$LRy.full)
}

#----------------------------------------------------------- Calculates decent direction
calc.h<-function(const,b,t.c,set,th){
  b.s<-as.matrix(b[set])
  ls.bs<-const$LRL.inv%*%const$LRy
  temp<-(t(th)%*%ls.bs-t.c)/(t(th)%*%const$LRL.inv%*%th)
  mu<-max(c(0,temp))
  h<-ls.bs-b.s-const$LRL.inv%*%th*mu
  list(h=h,lambda=mu)
}
```

```
#-------------------------------------------------------- Is new solution sign-feasible?
sign.feasible<-function(b,set,h.d,th){
  b.s<-b[set]
  len<-length(set)-1
  if (len==0)
    T
  else
    all(sign(b.s+h.d)[1:len]-sign(th)[1:len]==0)
}

#------------------------------------------------------------- Find most violated update
find.i<-function(b,set,h.d){
  b.s<-b[set]
  test<- -b.s/h.d
  maxi<-max(abs(test))
  len<-length(set)-1
  for (ii in 1:len)
    if (sign(b.s[ii])!=sign(b.s+h.d)[ii])
      if (test[ii]<=maxi){
        maxi<-test[ii]
        pos<-ii
      }
  pos
}

#------------------------------------------------------------ Calculate update step size
find.gamma<-function(vi,b,set,h.d){
  b.s<-b[set]
  -b.s[vi]/h.d[vi]
}

#---------------------------------------------------- Update predictions when downdating
update.b<-function(g,b,set,h.d){
  b.s<-b[set]
  b.s<-b.s+g*h.d
  b[set]<-b.s
  b
}

#------------------------------------------------------------------ Move to new solution
move<-function(b,h.d,set){
  b.s<-b[set]
  b.s<-b.s+h.d
  b[set]<-b.s
  b
}

#------------------------------------------------------------------ Finding non-zero set
find.set<-function(b){
  set<-double()
  kount<-1
  for (ii in 1:length(b))
    if (round(b[ii],10)!=0){
      set[kount]<-ii
      kount<-kount+1
    }
  if (kount!=1)
    set
  else
    NULL
}

#--------------------------------------- Reorder the iteration quantities for outputting
reorder.const<-function(const,set){
  if (length(set)<=1)
    val<-const
  else{
    LRL.inv<-const$LRL.inv[order(set),order(set)]
    LRL<-const$LRL[order(set),order(set)]
    LRy<-const$LRy[order(set)]
```

```
      val<-list(LRL.inv=LRL.inv,LRL=LRL,LRy=LRy,LRL.full=const$LRL.full,LRy.full=const$LRy.full)
    }
    class(val)<-"my.const"
    val
  }

#------------------------------------------------------------------------------------------
#                                 The lasso.t Function
#------------------------------------------------------------------------------------------

  if (!is.null(fm.prev))
    if (length(fm.prev$set)==0)
      fm.prev<-NULL
  iter.const<-init.const(LPL,LPy,fm.prev)
  new.S<-c()
  if (is.null(fm.prev))
    beta<-rep(0,length=length(LPy))
  else{
    beta<-fm.prev$coef
    new.S<-fm.prev$set
  }
  v<-calc.v(iter.const,beta,new.S)
  kount<-1
  kounter<-1
  while ( !is.optimal(v,new.S) | (kount==1) & (t!=0) ) {
    old.S<-new.S
    if ( (kount!=1) | is.null(fm.prev) ){
      new.S<-get.regressor(v,old.S)
      iter.const<-update.const(iter.const,new.S,old.S)
    }
    theta<-get.sign(beta,new.S,v)
    h<-calc.h(iter.const,beta,t,new.S,theta)
    while ((!sign.feasible(beta,new.S,h$h,theta))&(kounter<20)){
      old.S<-new.S
      vio<-as.integer(find.i(beta,new.S,h$h))
      gamma<-find.gamma(vio,beta,new.S,h$h)
      beta<-update.b(gamma,beta,new.S,h$h)
      new.S<-setdiff(new.S,new.S[vio])
      theta<-get.sign(beta,new.S,v)
      iter.const<-update.const(iter.const,new.S,old.S)
      h<-calc.h(iter.const,beta,t,new.S,theta)
      kount<-kount+1
      kounter<-kounter+1
      if (print.stuff==T) print(c("Sign Change Step,violated variable number: ",new.S[vio]))
    }
    beta<-move(beta,h$h,new.S)
    lam<-h$lambda
    v<-calc.v(iter.const,beta,new.S)
    kount<-kount+1
  }
  if ((kounter==20)&(print.stuff==T))
    print("Problem with sign change step")
  if (t==0){
    temp.S<-get.regressor(v,new.S)
    temp.const<-update.const(iter.const,temp.S,new.S)
    new.S<-NA
    theta<-get.sign(beta,temp.S,v)
    h<-calc.h(temp.const,beta,t,temp.S,theta)
    lam<-h$lambda
  }
  my.const<-reorder.const(iter.const,new.S)
  new.S<-sort(new.S)
  res<-list(coef=beta,v=v,lambda=lam,const=my.const,set=new.S)
  class(res)<-"lasso.t"
  res
}
```

4.2 Fitting the Penalised Regression for Correlated Observation

See Chapter 6 for details.

```
#---------------------------------------------------- Find t for given lambda by approx NR
lasso.lambda<-function(LPL,LPy,lambda.goal,delta.t=0.000001,maxit=50,tolerance=0.001,s.size=1,print.stuff=T)
#-----------------------------------------------------------------------------------------
#Calculates LASSO penalised regression predictions
#L is the LASSO design matrix, P is variance matrix of outcomes LPL=t(L)%*%P%*%L, y is the
#vector of outcomes, LPy=t(L)%*%P%*%y, lambda.goal is the given lambda to find t for,
#delta.t is the step size for t when numerically evaluating the differential, maxit is the
#max number of iterations allowed, tolerance is the desired precision and s.size is the
#step size to use when updating.

#Returns a list containing: t, lasso predictions, v, a boolean for converged, number of
#iterations, the iteration quantities, the active set of variables, lambda.goal and the
#maxium lambda for that data
#-----------------------------------------------------------------------------------------
{

  #---------------------------------------------------------------------------- Converged?
  converged<-function(tol,pt1,pt2)
    all(abs(pt1-pt2)<=tol)

#-----------------------------------------------------------------------------------------
#                             The lasso.lambda Function
#-----------------------------------------------------------------------------------------
  t.p<-h.dash<-double(length=maxit)
  h<-lambda<-matrix(ncol=2,nrow=maxit)

  ii<-1
  t.p[ii]<-0
  fm.1<-backup<-lasso.t(LPL=LPL,LPy=LPy,t=t.p[ii])
  lambda.max<-lambda[ii,1]<-fm.1$lambda
  if (length(lambda.goal)==0)
    lambda.goal<-fm.1$lambda+100
  if (lambda.goal > fm.1$lambda){
    lambda.goal<-fm.1$lambda
    too.big<-T
    if (print.stuff)
      print("Desired lambda too large -- assigned to largest value")
  }
  else{
    too.big<-F
    while ( (!converged(tol=tolerance,lambda.goal,lambda[ii])) & (ii<maxit)){
      fm.2<-lasso.t(LPL=LPL,LPy=LPy,t=(t.p[ii]+delta.t),fm.prev=fm.1)
      lambda[ii,2]<-fm.2$lambda
      h[ii,1]<-lambda[ii,1]-lambda.goal
      h[ii,2]<-lambda[ii,2]-lambda.goal
      h.dash[ii]<-abs(h[ii,2]-h[ii,1])/delta.t
      t.p[ii+1]<-t.p[ii]+s.size*(h[ii,1]/h.dash[ii])
      if (t.p[ii+1]<0){
        t.p[ii+1]<-t.p[ii]      #was 0.1*t.p[ii]
        print("In NR routine: updated t negative, this iteration assigned to zero")
      }
      if (t.p[ii+1]>t.p[ii])
        fm.1<-lasso.t(LPL=LPL,LPy=LPy,t=t.p[ii+1],fm.prev=fm.2)
      else
        fm.1<-lasso.t(LPL=LPL,LPy=LPy,t=t.p[ii+1],fm.prev=backup)
      if (fm.1$lambda>lambda.goal)
        backup<-fm.1
      lambda[ii+1,1]<-fm.1$lambda
      ii<-ii+1
    }
  }
  if (ii==maxit){
    conv<-F
    print("convergence failed in finding t")
  }
  else
```

```
    conv<-T
  val<-list(t=t.p[ii],coef=fm.1$coef,v=fm.1$v,converged=conv,niter=ii,const=fm.1$const,
                    set=fm.1$set,lambda=fm.1$lambda,lambda.max=lambda.max,too.big=too.big)
  val
}
```

4.3 Solving the LASSO Mixed Model Equations

See Chapter 8 for details.

```
#----------------------------------------------------------------------------- Solves LMME
gauss.elim<-function(C,lambda,p,q,r,print.stuff=T,fm=NULL)
#-----------------------------------------------------------------------------------------
#Solves the LMME by modified Gaussian elimination
#C is the LMME matrix appended with the 'modified outcome' vectors, lambda is the penalty
#for the LASSO equations, p is #fixed effects, q is #LASSO effects, r is a vector with
#each element describing the number of random effects for that group, print.stuff is a
#boolean for trace printing and fm is a previous lasso.lambda model for starting values.

#Function returns a list: a list of fixed, LASSO and random estimates/predictions, the
#adjusted sums of squares, the l1 norm of LASSO effects, v, the set of active LASSO vars,
#various quantities for later use.
#-----------------------------------------------------------------------------------------
{
  if ((p!=0)|(sum(r)!=0))
    A<-absorb(mat=C,t=1,q=q,p=p,r=r,to="lasso",log.det=FALSE)
  else
    A<-list(absorbed=C,log.det=NULL)

  if (is.null(fm)){
    fm<-lasso.lambda(LPL=A$absorbed[1+1:q,1+1:q],LPy=A$absorbed[1,1+1:q],
             lambda.goal=lambda,tolerance=1e-10,maxit=30,s.size=1,print.stuff=print.stuff)
    if (fm$too.big)
      lambda<-fm$lambda.max
    beta<-fm$coef
  }
  else
    beta<-fm$coef$lasso
  esti<-b.solve(mat=A$absorbed,beta=beta,p=p,q=q,r=r)
  RSS<-A$absorbed[1,1]-t(beta)%*%A$absorbed[1+1:q,1+1:q]%*%beta
  vb<-as.double(t(fm$v)%*%beta)
  v<-fm$v
  set<-fm$set
  LPL.invS<-fm$const$LRL.inv

  res<-list(coef=esti,RSS=RSS,vb=vb,v=v,set=set,LPL.invS=LPL.invS,
                               LPL=A$absorbed[1+1:q,1+1:q],LPy=A$absorbed[1+1:q,1],t=fm$t)
  res
}


#------------------------------------------------------------------------ Absorbs a matrix
absorb<-function(mat,t=1,q,p,r,to="lasso",log.det=FALSE)
#-----------------------------------------------------------------------------------------
#mat is the matrix which will be absorbed to either the block of equations to="top" or
#to="lasso", t is the dim of the appended equations, q is the no. of the LASSO vars, p is
#the no. of fixed effects and r is the vector of numbers of random effects, log.det==TRUE
#stipulates that the log det will be calculated, assumes that fixed and ranodm normal
#effects will be absorbed.

#Function returns a list: the absorbed matrix, log-det of the absorbed equations.
#-----------------------------------------------------------------------------------------
{
  new.mat<-mat
  #absorb the random effects
  if (sum(r)!=0)
    for (ii in 1:length(r)){
      if (length(r)!=ii){
        dim.remain<-1:(t+q+p+sum(r[1:(length(r)-ii)]))
        dim.spent<-setdiff(1:(t+q+p+sum(r[1:(length(r)-ii+1)])),dim.remain)
      }
      else{
        dim.remain<-1:(t+q+p+0)
        dim.spent<-setdiff(1:(t+q+p+r[1]),dim.remain)
      }
      new.mat[dim.remain,dim.remain]<-new.mat[dim.remain,dim.remain]-new.mat[dim.remain,dim.spent]%*%
                                        solve(new.mat[dim.spent,dim.spent])%*%new.mat[dim.spent,dim.remain]
      new.mat[dim.remain,dim.spent]<-0
    }
  if (p!=0){
    dim.remain<-1:(t+q)
    dim.spent<-(t+q+1):(t+q+p)
    new.mat[dim.remain,dim.remain]<-new.mat[dim.remain,dim.remain]-new.mat[dim.remain,dim.spent]%*%
                                        solve(new.mat[dim.spent,dim.spent])%*%new.mat[dim.spent,dim.remain]
    new.mat[dim.remain,dim.spent]<-0
  }
  if (to=="top"){
    new.mat[1:t,1:t]<-new.mat[1:t,1:t]-new.mat[1:t,t+1:q]%*%solve(new.mat[t+1:q,t+1:q])%*%new.mat[t+1:q,1:t]
    new.mat[1:t,t+1:q]<-0
  }


  if (log.det)
    mat.det<-switch(to,
      "lasso"=log(det(as.matrix(mat[t+q+1:(rp.sum),t+q+1:rp.sum],ncol=rp.sum))),
      "top"=log(det(as.matrix(mat[t+1:(q+rp.sum),t+1:(q+rp.sum)],ncol=q+rp.sum))))
  else
    mat.det<-NULL

  val<-list("absorbed"=new.mat,"log.det"=mat.det)
  val
}


#--------------------------------------------------- Backsolves a previously absorbed LMME
b.solve<-function(mat,beta,p,r,q)
#-----------------------------------------------------------------------------------------
#mat is an absorbed matrix to the lasso effects, the LMME are assumed to have only the
#outcomes appended. beta is a vector of the estimated lasso effects, p is the no. of fixed
#effects, r is the vector containing the no. of random effects and q is the number of
#lasso effects.

#Function returns a list of the vectors: 'fixed', 'lasso' and 'random' corresponding to
#the respective effects

#-----------------------------------------------------------------------------------------
{
  if (is.null(beta))
    print("beta must be specified for back solving routine, function will fall over")
  rp.sum<-p+sum(r)
  solved<-matrix(nrow=q+rp.sum,ncol=1)
  solved[1:q,1]<-beta
  if (p!=0)
    fixed<-solved[q+1:p,1]<-solve(mat[1+q+1:p,1+q+1:p])%*%(mat[1+q+1:p,1]-mat[1+q+1:p,1+1:q]%*%solved[1:q])
  else
    fixed<-c()
  if (sum(r)!=0){
    random<-list()
    for (ii in 1:length(r)){
      if (ii!=1)
        current<-1+q+p+sum(r[1:(ii-1)])+1:r[ii]
      else
        current<-1+q+p+1:r[ii]
      current.effect<-intersect(1:(1+q+p+sum(r)),current)-1
      previous<-2:(min(current)-1)
      previous.effect<-1:(min(current)-2)
      random[[ii]]<-solved[current.effect,1]<-solve(mat[current,current])%*%(mat[current,1]
                                                   -mat[current,previous]%*%solved[previous.effect])
    }
  }
  else
    random<-NULL

  val<-list(fixed=as.double(fixed),lasso=as.double(beta),random=random)
  val
}
```

4.4 Fitting the LASSO Random Effects Model and the LASSO Linear Mixed Model (LLMM)

See Chapter 7 for details on the random LASSO effect only model. See Chapter 8 for details on the LLMM. The function `LLMM.boot()` calls `update.comps.boot()` and will be the one that the user calls.

Updating Components

```
#---------------------------------------------------------- Calc scores for all gammas
get.gammas.scores<-function(sigma2,G.i,fm,r,q,p,n.random,m.list.here=NULL)
#-------------------------------------------------------------------------------------
#Slave function to LLMM.boot and update.comps.boot
#function returns the (potentially bootstrapped adjusted) scores for gamma
{

  score.gamma<-function(sigma2,G.dot,G.inv,u.tilde,r)
    t(u.tilde)%*%G.inv%*%G.dot%*%G.inv%*%u.tilde/(2*sigma2)

  scores<-double(length=n.random)
  before<-0
  for (jj in 1:n.random){
    temp.G.inv<-G.i$inv[before+1:r[jj],before+1:r[jj]]
    scores[jj]<-score.gamma(sigma2=sigma2,G.dot=G.i$dot[[jj]],G.inv=temp.G.inv,
                                                  u.tilde=fm$coef$random[[jj]],r=r[jj])
    before<-before+r[jj]
  }
  if (is.null(m.list.here))
    scores
  else
    scores-m.list.here$gamma/sigma2
}

#-------------------------------------------------------- Updates disperion components
update.comps.boot<-function(fm,prev,n,p,q,kappa,r=r,s.s,max.z,n.boot=100,C.star,quads,
                                       m.list,n.random,G.i,esti.sig.phi,type=type,kount1)
#-------------------------------------------------------------------------------------
{
  #------------------------------------------------ Update estimates for sigma and phi
  get.sig.phi<-function(fm,prev,n,q,kappa,p,r,n.random,n.boot,G.i,m.list,s.s,max.z,type)
  {
    q.s<-length(fm$set)
    if (q.s!=0){
      temp.RSS<-as.double(fm$RSS-2*prev$sigma*fm$vb/prev$phi)
      if (type=="boot"){
        new.phi<-fm$vb/(q+prev$phi*m.list$phi)
        new.sigma<-as.double(temp.RSS/(n-p-kappa+2*prev$sigma*m.list$sigma))
      }
      else{
        new.phi<-fm$vb/q.s
        new.sigma<-as.double(temp.RSS/(n-p-q.s))
      }
      new.phi<-as.double(prev$phi+s.s*(new.phi-prev$phi))
      new.sigma<-as.double(prev$sigma+s.s*(new.sigma-prev$sigma))
      new.z.kount<-prev$z.kount
      able<-TRUE
      step.flag<-FALSE
    }
    else{
      temp.RSS<-as.double(fm$RSS)
      if (prev$z.kount<max.z){
        print("Next iterate of phi zero, step.size reduced by 25%")
        new.z.kount<-prev$z.kount+1
        temp.s.s<-(0.25^new.z.kount)*s.s
        new.phi<-temp.s.s*prev$phi
        if (type=="boot")
          new.sigma<-prev$sigma+
                 s.s*(as.double(temp.RSS/(n-p-kappa+2*prev$sigma*m.list$sigma)-prev$sigma))
        else
          new.sigma<-prev$sigma+s.s*(as.double(temp.RSS/(n-p-q.s)-prev$sigma))
        able<-FALSE
        step.flag<-FALSE
      }
      else{
        m.list<-list(sigma=0,phi=0)
        new.phi<-0
        new.sigma<-temp.RSS/(n-p)
        new.z.kount<-prev$z.kount
        able<-TRUE
        step.flag<-TRUE
      }
    }
    list(sigma=new.sigma,phi=new.phi,z.kount=new.z.kount,able=able,step.flag=step.flag)
  }

  #--------------------------------------------------------------- Calc average information
  get.info<-function(g.set,n,q,p,quads,sig,n.random,m.list)
  {
    n.g<-length(g.set)
    info<-matrix(0,ncol=n.g+1,nrow=n.g+1)
    info[1,1]<-quads$q1Pq1/(2*sig)
    if (!is.null(g.set))
      for (ii in 1+1:n.g){
        info[1,ii]<-info[ii,1]<-quads$q1Pqi[g.set[ii-1]]/(2*sig^2)-m.list$gamma[ii-1]/(sig^2)
        if (n.random!=1)
          for (jj in 1+1:n.g)
            info[ii,jj]<-info[jj,ii]<-quads$qiTqj[g.set[ii-1],g.set[jj-1]]/sig
        else
          info[2,2]<-quads$qiTqj/sig
      }
    info.inv<-solve(info)
    list(info=info,info.inv=info.inv)
  }

  #----------------------------------------------------------------------- moves the gammas
  update.gamma<-function(g.set,prev.gamma,s.s,info.inv,scores,print.stuff=T)
  {
    n.g<-length(g.set)
    new.gamma<-double(length=length(prev.gamma))
    if (!is.null(g.set)){
      #### negative or positive?
      new.gamma[g.set]<-prev.gamma[g.set]+s.s[2]*info.inv[1+1:n.g,1+1:n.g]%*%scores[g.set]
      not.g.set<-setdiff(1:length(prev.gamma),g.set)
      new.gamma[not.g.set]<-prev.gamma[not.g.set]
      g.set<-NULL
      for (ii in 1:n.g){
        if (new.gamma[ii]<0){
          if (prev.gamma[ii]>0.00001){
            new.gamma[ii]<-0.1*prev.gamma[ii]
            if (print.stuff==T)
              print("Updated gamma(s) negative, reduced by 10% and removed from descent step")
          }
          else{
            new.gamma[ii]<-0.00001
            if (print.stuff==T)
              print("Updated gamma(s) negative and gamma small. Assigned to 0.00001")
          }
        }
        else{
          if (!is.element(ii, not.g.set))
            g.set<-c(g.set,ii)
        }
      }
    }
    else
      new.gamma<-prev.gamma
    list(new.gamma=new.gamma,g.set=g.set,prev.gamma=prev.gamma)
  }


#-----------------------------------------------------------------------------------------
#                           The Disperion updating function
#-----------------------------------------------------------------------------------------

  if (esti.sig.phi==TRUE)
    sig.phi<-get.sig.phi(fm=fm,prev=prev,n=n,q=q,kappa=kappa,p=p,r=r,n.random=n.random,
                          n.boot=n.boot,G.i=G.i,m.list=m.list,s.s=s.s[1],max.z=max.z,type)
  else
    sig.phi<-list(sigma=prev$sigma,phi=prev$phi,z.kount=prev$z.kount,
                                             able=prev$able,step.flag=prev$step.flag,type)
  if (prev$z.kount<max.z)
    C.star<-NULL
  if (n.random!=0){
    score.gamma<-get.gammas.scores(sigma2=prev$sigma,G.i=G.i,fm=fm,r=r,q=q,p=p,
                                                     n.random=n.random,m.list.here=m.list)
    if (kount1==1)
      get.set<-1:n.random
    else
      get.set<-prev$gamma.set
    active<-n.random+1
    while (length(get.set)<active){
      active<-length(get.set)
      av.info<-get.info(g.set=get.set,n=n,q=q,p=p,quads=quads,sig=prev$sigma,
                                                          n.random=n.random,m.list=m.list)
      new.gamma<-update.gamma(g.set=get.set,prev.gamma=prev$gamma,s.s=s.s,
                               info.inv=av.info$info.inv,scores=score.gamma,print.stuff=T)
      prev$gamma<-new.gamma$new.gamma
      get.set<-new.gamma$g.set
      if (is.null(get.set))
        active<-0
      score.gamma<-get.gammas.scores(sigma2=prev$sigma,G.i=G.i,fm=fm,r=r,q=q,p=p,
                                                     n.random=n.random,m.list.here=m.list)
    }
  }
  else
    new.gamma<-score.gamma<-get.set<-c()

  list(sigma=as.double(sig.phi$sigma),phi=as.double(sig.phi$phi),
      gamma=new.gamma$new.gamma,score.gamma=score.gamma,gamma.set=get.set,
          z.kount=sig.phi$z.kount,able.conv=sig.phi$able,step.flag=sig.phi$step.flag,
              M=c(2*prev$sigma*m.list$sigma,prev$phi*m.list$phi))
}
```

The Governing LLMM Function

```
#-------------------------------------------------------------------------- Forms the LMME
get.C<-function(top,L,X,Z,G.inv,n,q,p,r,n.random)
#-----------------------------------------------------------------------------------------
#slave function to LLMM.boot and update.comps.boot
#Function returns the appended LMME
{
  t<-length(top[1,])
  K<-matrix(nrow=n,ncol=t+q+p+sum(r))
  if (p!=0)
    K[,1:(t+q+p)]<-cbind(top,L,X)
  else
    K[,1:(t+q)]<-cbind(top,L)
  if (n.random!=0){
    present<-t+q+p
    for (ii in 1:n.random){
      K[,present+1:r[ii]]<-Z[[ii]]
      present<-present+r[ii]
    }
    C<-t(K)%*%K
    C[t+q+p+1:sum(r),t+q+p+1:sum(r)]<-C[t+q+p+1:sum(r),t+q+p+1:sum(r)]+G.inv
  }
  else
    C<-t(K)%*%K
  C
}

#----------------------------------------------- Fits LLMM using bootstrap adjusted scores
LLMM.boot<-function(y,L,X,Z=NULL,n.random=0,inits=list(sigma=NULL,phi=NULL,gamma=NULL),maxit=c(100,10,10),
                               tolerance=c(0.001,0.01),step.size=c(1,0.95,1),n.boot=c(100,0,500),type="boot")
#-----------------------------------------------------------------------------------------
#y is outcome vector, L is design matrix for LASSO effects, X is design matrix for fixed
#effects, Z is a list of design matrices for random effects, n.random=length(Z), inits are
#initial dispersion parameters (if any), maxit[1] is the total number of parameter updates
#to allow, maxit[2] is the max no. of iterations to allow phi to go to zero before
#assigning it to zero, maxit[3] is the max number of updates to perform for each iteration
#of gamma, tolerance are tolerance levels for overall convergence and gamma convergence,
#step.size are for [1] parameter updates, [2] for reducing [1] after each iteration and [3]
#for updating gamma, n.boot [1] is the number of initial bootstrap samples, [2] increse
#amount after each iteration, [3] max no. of bootstrap samples and type="boot" for bootstrap
#estimation and "qs" for the ad-hoc adjustment.

#Function returns effects (fixed, lasso and random), sigma, phi, gamma, RSS, set of active
#LASSO effects, convergence boolean, v, fitted values, residuals, t and some estimation quantities.
{

  #----------------------------------------------------------------- Update the G matrices
  update.G<-function(gamma,n.random,r)
  {
    if (n.random!=0){
      G.dot<-list(length=n.random)
      G.diag<-double(legnth=sum(r))
      present<-1
      for (ii in 1:n.random){
        G.dot[[ii]]<-diag(rep(1,r[ii]))
        G.diag[present:sum(r[1:n.random<=ii])]<-gamma[ii]
        present<-present+r[ii]
      }
      G<-diag(G.diag)
      G.inv<-diag(1/G.diag)
      list(G=G,inv=G.inv,dot=G.dot)
    }
    else
      list(G=NULL,inv=NULL,dot=NULL)
  }


  #--------------------------------------------------------------------- Check convergence

  converged<-function(tol,pt1,pt2,able=T)
  {
    if (able){
      if (all(abs(pt1-pt2)<=tol))
        "TRUE"
      else
        "FALSE"
    }
    else
      FALSE
  }
```

```
#----------------------------------------------------- Check convergence of gamma scores
converged.gamma<-function(tol,pt1,pt2,g.set)
{
  if (!is.null(g.set)){
    if (all(abs(pt1[g.set]-pt2[g.set])<=tol))
      TRUE
    else
      FALSE
  }
  else
    TRUE
}


#---------------------------------------------------------- Initialises model parameters
init.model<-function(C,q,p,r,inits)
{
  if ((sum(r)+p)!=0)
    A<-absorb(mat=C,t=1,q=q,p=p,r=r,to="lasso",log.det=FALSE)
  else
    A<-list(absorbed=C,log.det=NULL)
  fm<-lasso.t(LPL=A$absorbed[1+1:q,1+1:q],LPy=A$absorbed[1+1:q,1],t=0)
  lambda.max<-fm$lambda
  if ((is.null(inits$sigma)) & (is.null(inits$phi)))
    start.lambda<-0.1*lambda.max
  else{
    start.lambda<-inits$sigma/(inits$phi)
    if (start.lambda>lambda.max)
      start.lambda<-0.1*lambda.max
  }
  list(lambda.max=lambda.max,start.lambda=start.lambda)

}

#----------------------------------------------------------- Initialises disp parameters
init.comp<-function(fm,inits,n,p,r,max.lambda,n.random)
{
  q.s<-length(fm$set)
  if (is.null(inits$phi))
    new.phi<-as.double(fm$vb/q.s)
  else
    new.phi<-inits$phi
  if (is.null(inits$sigma))
    new.sigma<-as.double(fm$RSS/(n-p-q.s+2*fm$vb/new.phi))
  else
    new.sigma<-inits$sigma
  while ((new.sigma/new.phi)>=max.lambda){
    new.phi<-2*new.phi
    new.sigma<-0.5*new.sigma
  }
  z.kount<-0
  if (n.random!=0)
    gamma.set<-1:n.random
  else
    gamma.set<-NULL

  list(sigma=new.sigma,phi=new.phi,gamma=inits$gamma,score.gamma=rep(-99,length(r)),
   able.conv=T,step.flag=NULL,M=c(NA,NA),z.kount=z.kount,gamma.set=gamma.set)
}

#------------------------------------------------------------------ Bootstrapping scores
get.bias<-function(n,p,q,kappa,r,n.random,B,prev,L,X,Z,fm,G.i)
{
  score.sigma<-function(n,p,q,kappa,sigma2,phi,RSS,vb)
    as.double(-(n-p-kappa)/(2*sigma2)+RSS/(2*(sigma2^2))-vb/(sigma2*phi))
  score.phi<-function(q,phi,vb)
    as.double(-q/phi+vb/(phi^2))

  score.boot<-matrix(ncol=2+n.random,nrow=B)
  max.int<-ceiling(B/50)
  if (n.random==0){
    dimnames(score.boot)[[2]]<-c("sigma","phi")
    for (ii in 1:B){
      b.star<-matrix((2*rbinom(n=length(L[1,]),size=1,prob=0.5)-1)*rexp(n=length(L[1,]),
                                                                  rate=1/prev$phi),ncol=1)
      e.star<-matrix(rnorm(n=n,mean=0,sd=sqrt(prev$sigma)),ncol=1)
      if (p!=0)
        y.star<-X%*%fm$coef$fixed+L%*%b.star+e.star
      else
        y.star<-L%*%b.star+e.star
      C.star<-get.C(top=y.star,L=L,X=X,Z=Z,G.inv=G.i$inv,n=n,q=q,p=p,r=r,n.random=n.random)
      if (prev$phi!=0){
        fm1<-gauss.elim(C=C.star,lambda=prev$sigma/prev$phi,p=p,q=q,r=r,print.stuff=F)
        score.boot[ii,1]<-score.sigma(n=n,p=p,q=q,kappa=kappa,sigma2=prev$sigma,
                                                     phi=prev$phi,RSS=fm1$RSS,vb=fm1$vb)
        score.boot[ii,2]<-score.phi(q=q,phi=prev$phi,vb=fm1$vb)
      }
      else{
        fm1<-gauss.elim(C=C.star,lambda=prev$sigma/0.0000001,p=p,q=q,r=r,print.stuff=F)
        score.boot[ii,1]<-score.sigma(n=n,p=p,q=q,kappa=kappa,sigma2=prev$sigma,
                                                            phi=1,RSS=fm1$RSS,vb=fm1$vb)
        score.boot[ii,2]<- 0
      }
      if (is.element(ii/50,1:max.int))
        print(c("bootstrap iteration",ii))
    }
    gamma.res<-NULL
  }
  else{
    dimnames(score.boot)[[2]]<-c("sigma","phi",paste("gamma",1:n.random,sep="_"))
    for (ii in 1:B){
      Zu.star<-matrix(rep(0,n),ncol=1)
      for (jj in 1:n.random)
        Zu.star<-Zu.star+Z[[jj]]%*%matrix(rnorm(n=r[jj],mean=0,
                                             sd=sqrt(prev$sigma*prev$gamma[jj])),ncol=1)
      b.star<-matrix((2*rbinom(n=length(L[1,]),
                      size=1,prob=0.5)-1)*rexp(n=length(L[1,]),rate=1/prev$phi),ncol=1)
      e.star<-matrix(rnorm(n=n,mean=0,sd=sqrt(prev$sigma)),ncol=1)
      if (p!=0)
        y.star<-X%*%fm$coef$fixed+Zu.star+L%*%b.star+e.star
      else
        y.star<-Zu.star+L%*%b.star+e.star
      C.star<-get.C(top=y.star,L=L,X=X,Z=Z,G.inv=G.i$inv,n=n,q=q,p=p,r=r,n.random=n.random)
      if (prev$phi!=0){
        fm1<-gauss.elim(C=C.star,lambda=prev$sigma/prev$phi,p=p,q=q,r=r,print.stuff=F)
        score.boot[ii,1]<-score.sigma(n=n,p=p,q=q,kappa=kappa,sigma2=prev$sigma,
                                                     phi=prev$phi,RSS=fm1$RSS,vb=fm1$vb)
        score.boot[ii,2]<-score.phi(q=q,phi=prev$phi,vb=fm1$vb)
      }
      else{
        fm1<-gauss.elim(C=C.star,lambda=prev$sigma/0.0000001,p=p,q=q,r=r,print.stuff=F)
        score.boot[ii,1]<- score.sigma(n=n,p=p,q=q,kappa=kappa,sigma2=prev$sigma,
                                                            phi=1,RSS=fm1$RSS,vb=fm1$vb)
        score.boot[ii,2]<- 0
      }
      score.boot[ii,2+1:n.random]<-get.gammas.scores(sigma2=1,G.i=G.i,fm=fm1,
                                         r=r,q=q,p=p,n.random=n.random,m.list.here=NULL)
      if (is.element(ii/50,1:max.int))
        print(c("bootstrap iteration",ii))
    }
    gamma.res<-double(length=n.random)
    for (jj in 1:n.random)
      gamma.res[jj]<-mean(score.boot[,2+jj],trim=0.05)
  }


  list(sigma=mean(score.boot[,1],trim=0.05),phi=mean(score.boot[,2],trim=0.05),gamma=gamma.res)
}
```

```
  #-------------------------------------------- Calculate working variables for absorption
  get.working<-function(y,fm,L,Z,r,n.random,G)
  {
    beta<-fm$coef$lasso
    u<-fm$coef$random

    q1<-y-L%*%beta
    qi<-matrix(ncol=n.random,nrow=length(L[,1]))
    place<-0
    for (ii in 1:n.random){
      qi[,ii]<-Z[[ii]]%*%G$dot[[ii]]%*%G$inv[place+1:r[ii],place+1:r[ii]]%*%u[[ii]]
      place<-place+r[ii]
    }
    cbind(q1,qi)
  }

  #--------------------------------------------------------- Obtain quadratic forms for AI
  get.quads<-function(C,t,q,p,r,n.random,set)
  {
    A1<-absorb(mat=C,t=t,q=q,p=p,r=r,to="lasso")
    q.s<-length(set)
    if (q.s!=0)
      A2<-absorb(mat=A1$absorbed[c(1:t,t+set),c(1:t,t+set)],t=t,q=q.s,p=0,r=0,to="top")
    else
      A2<-list(absorbed=A1$absorbed[1:t,1:t])
    list(q1Pq1=A1$absorbed[1,1],q1Pqi=A1$absorbed[1,1+1:n.random],
                                             qiTqj=A2$absorbed[1+1:n.random,1+1:n.random])
  }



#-----------------------------------------------------------------------------------------
#                                The LLMM.boot Function
#-----------------------------------------------------------------------------------------
  n<-length(y)
  q<-length(L[1,])
  if (!is.null(X))
    p<-length(X[1,])
  else
    p<-0
  if (n.random!=0){
    r<-double(length=n.random)
    for (ii in 1:n.random)
      r[ii]<-length(Z[[ii]][1,])
  }
  else
    r<-0
  s.s.orig<-step.size[1]
  if (is.null(inits$gamma))
    inits$gamma<-rep(0.1,n.random)
  G.i<-update.G(gamma=inits$gamma,n.random,r=r)
  C.star<-get.C(top=y,L=L,X=X,Z=Z,G.inv=G.i$inv,n=n,q=q,p=p,r=r,n.random=n.random)
  dum<-init.model(C=C.star,q=q,p=p,r=r,inits=inits)
  fm<-gauss.elim(C=C.star,lambda=dum$start.lambda,p=p,q=q,r=r)
  var.comp<-init.comp(fm=fm,inits=inits,n=n,p=p,r=r,max.lambda=dum$lambda.max,n.random=n.random)
  flag<-kount<-1
  prev.comp<-list(sigma=-99,phi=-99,gamma=rep(-99,n.random),z.kount=0)

  if (n.random!=0)
    print(c("Iteration","Sigma2","Phi",paste("Gamma",1:n.random,sep="_"),"q.s","M.sigma","M.phi"))
  else
    print(c("Iteration","Sigma2","Phi","q.s","M.sigma","M.phi"))

  conv<-F

  while ((!converged(pt1=c(var.comp$sigma,var.comp$phi,var.comp$gamma),
               pt2=c(prev.comp$sigma,prev.comp$phi,prev.comp$gamma),tol=tolerance[1],able=var.comp$able))
                                                  & (kount<=maxit[1])){
    kount1<-1
    prev.comp<-var.comp
```

```
e.vals<-eigen(C.star[1+1:q,1+1:q])$values
kappa<-length(e.vals[round(e.vals,5)>0])
if (kount!=1){
  if (type=="boot"){
    print("Bootstrapping for bias correction")
    m.list<-get.bias(n=n,p=p,q=q,kappa=kappa,r=r,n.random=n.random,B=n.boot[1],
                                              prev=var.comp,L=L,X=X,Z=Z,fm=fm,G.i=G.i)
    print("Done bootstrapping")
  }
  else
    m.list<-list(sigma=0,phi=0,gamma=rep(0,n.random))
  if (n.random!=0){
    get.sigma.phi<-TRUE
    var.comp$gamma.set<-1:n.random
    while ((!converged.gamma(pt1=var.comp$score.gamma,pt2=rep(0,n.random),
                      g.set=var.comp$gamma.set,tol=tolerance[2]))&(kount1<=maxit[3])){
      Q<-get.working(y=y,fm=fm,L=L,Z=Z,r=r,n.random=n.random,G=G.i)
      C.starS<-get.C(top=Q,L=L,X=X,Z=Z,G.inv=G.i$inv,n=n,q=q,p=p,r=r,n.random=n.random)
      quads<-get.quads(C=C.starS,t=length(Q[1,]),q=q,p=p,r=r,n.random=n.random,set=fm$set)
      var.comp<-update.comps.boot(fm=fm,prev=var.comp,n=n,p=p,q=q,kappa=kappa,
                    s.s=step.size[c(1,3)],max.z=maxit[2],C.star=C.star,quads=quads,
                        m.list=m.list,n.random=n.random,G.i=G.i,r=r,esti.sig.phi=get.sigma.phi,
                                                                          type=type,kount1=kount1)
      get.sigma.phi<-FALSE
      G.i<-update.G(gamma=var.comp$gamma,n.random,r=r)
      C.star<-get.C(top=y,L=L,X=X,Z=Z,G.inv=G.i$inv,n=n,q=q,p=p,r=r,n.random=n.random)
      fm<-gauss.elim(C=C.star,lambda=NULL,p=p,q=q,r=r,fm=fm)
      kount1<-kount1+1
    }
    if (kount1>maxit[3])
      print("gamma update not converged!")
    var.comp$gamma[var.comp$gamma.set]<-prev.comp$gamma[var.comp$gamma.set]+
            step.size[1]*(var.comp$gamma[var.comp$gamma.set]-prev.comp$gamma[var.comp$gamma.set])
    var.comp$score.gamma<-rep(-9999,n.random)
  }
  else
    var.comp<-update.comps.boot(fm=fm,prev=var.comp,n=n,q=q,kappa=kappa,p=p,
               s.s=step.size[c(1,3)],max.z=maxit[2],C.star=C.star,quads=NULL,m.list=m.list,
                             n.random=n.random,G.i=G.i,r=r,esti.sig.phi=TRUE,type=type,kount1=1)
}
else{
  if (n.random==0)
    prev.comp<-list(sigma=-99,phi=-99,gamma=NULL)
  if (n.random!=0){
    m.list<-list(gamma=rep(NA,n.random))
    prev.comp<-list(sigma=-99,phi=-99,gamma=rep(-99,n.random))
  }
}
if ((is.na(var.comp$phi))|(var.comp$phi==0)){
  var.comp$phi<-0
  dum<-init.model(C=C.star,q=q,p=p,r=r,inits=inits)
  fm<-gauss.elim(C=C.star,lambda=dum$lambda.max,p=p,q=q,r=r,fm=NULL)
  if (var.comp$step.flag){
    print("Phi identically zero, step.size assigned to 1")
    step.size[1]<-1
    var.comp$step.flag<-FALSE
  }
}
else
  fm<-gauss.elim(C=C.star,lambda=var.comp$sigma/var.comp$phi,p=p,q=q,r=r,fm=NULL)

if (n.random!=0)
    print(c(kount,round(var.comp$sigma,4),round(var.comp$phi,4),
         round(var.comp$gamma,4),length(fm$set),round(var.comp$M[1],4),round(var.comp$M[2],4)))
else
    print(c(kount,round(var.comp$sigma,4),round(var.comp$phi,4),length(fm$set),
                                                round(var.comp$M[1],4),round(var.comp$M[2],4)))

if ((kount>1)&(var.comp$phi!=0)){
  step.size[1]<-step.size[1]*step.size[2]
```

```
      print(c("step size reduced now:",round(step.size[1],3)))
      if (type=="boot"){
        if (n.boot[1]<n.boot[3]){
          n.boot[1]<-n.boot[1]+n.boot[2]
          print(c("Number of bootstrap samples increased, now:", n.boot[1]))
        }
        else
          print("Number of bootstrap samples not increased, already at largest limit")
      }
    }

    kount<-kount+1

  }

  if ((kount-1)==maxit[1]) conv<-F else conv<-T
  if (!conv) print("Failed to Converge") else print("Converged")

  if (is.null(var.comp$phi))
    var.comp$phi<-0

  fitted<-L%*%fm$coef$lasso
  if (!is.null(X))
    fitted<-X%*%fm$coef$fixed+fitted
  if (n.random!=0)
    for (ii in 1:n.random)
      fitted<-fitted+Z[[ii]]%*%fm$coef$random[[ii]]
  residual<-y-fitted

  res<-list(fixed=fm$coef$fixed,lasso=fm$coef$lasso,random=fm$coef$random,
          sigma=var.comp$sigma,phi=var.comp$phi,gamma=var.comp$gamma,RSS=var.comp$RSS,
              set=fm$set,conv=conv,v=fm$v,fitted=fitted,residuals=residual,LPL=fm$LPL,LPy=fm$LPy,t=fm$t)
  class(res)<-"LLMM.boot"
  res
}
```

4.5 Hypothesis Testing for LASSO Predictions

See Chapters 7 and 8 for details.

```
#------------------------------------------------- Simulation Testing of LASSO Predictions
LASSO.sim.test<-function(obj,X=NULL,Z=NULL,L,B=100,alpha=c(0.01,0.05,0.1),plot.dist=FALSE,return.dist=FALSE)
#-----------------------------------------------------------------------------------------
#Finds critical values for testing the LASSO predictions (both experiment and comparison wise)
#obj is a Splus function fiited using LLMM.boot. X,Z and L are the fixed, random and LASSO
#effect design matrices (Z a list), B is the number of simulation samples, alpha is the
#vector of critical value probs required, plot.dist is a boolean (if TRUE then the dist.
#of the predictions is plotted) and return.dist is a boolean (if TRUE then the full set
#of sampled predictions (under null hypoth) are returned).

#Function returns a list containing: experiment-wise and comparison-wise critical values
#for the alpha level and if return.dist==T then the full set of simulated predictions.
{
  get.C.base<-function(L,X.mat,Z,G.inv,n,q,p,r,n.random){
    K<-matrix(nrow=n,ncol=q+p+sum(r))
    K[,1:(q+p)]<-cbind(L,X.mat)
    if (n.random!=0){
      present<-q+p
      for (ii in 1:n.random){
        K[,present+1:r[ii]]<-Z[[ii]]
        present<-present+r[ii]
      }
      C<-t(K)%*%K
      C[q+p+1:sum(r),q+p+1:sum(r)]<-C[q+p+1:sum(r),q+p+1:sum(r)]+G.inv
    }
    else
      C<-t(K)%*%K
    C
  }

  esti.fun<-function(X,X.mat,L,Z.mat,C.base,lambda,q,p,r,n.random){
    pqr<-p+q+sum(r)
    C.star<-matrix(ncol=pqr+1,nrow=pqr+1)
    C.star[1+1:pqr,1]<-C.star[1,1+1:pqr]<-as.double(t(X)%*%cbind(L,X.mat,Z.mat))
    C.star[1,1]<-t(X)%*%X
    C.star[-1,-1]<-C.base
    A.star<-absorb(mat=C.star,to="lasso",t=1,q=q,p=p,r=r)
    LPy.star<-A.star$absorbed[1+1:q,1]
    LPL.star<-A.star$absorbed[1+1:q,1+1:q]
    fm.star<-lasso.lambda(LPL=LPL.star,LPy=LPy.star,lambda.goal=lambda,print.stuff=F)
    temp<-rev(order(abs(fm.star$coef)))
    temp1<-c(temp[1],abs(fm.star$coef[temp[1]]),abs(fm.star$coef))
    temp1
  }

  n<-length(obj$fitted)
  a<-length(alpha)
  maxes<-matrix(ncol=a,nrow=B)
  q<-length(L[1,])
  if (!is.null(X)){
    mean.Xt<-as.double(X%*%obj$fixed)
    p<-length(X[1,])
  }
  else{
    p<-0
    mean.Xt<-rep(0,length(L[,1]))
  }
  if (is.null(Z)){
    r<-0
    H.mat<-obj$sigma*diag(1,n)
    n.random<-0
    Z.mat<-NULL
  }
  else{
    n.random<-length(Z)
    r<-double(length=n.random)
    H.mat<-obj$sigma*diag(1,n)
    G.diag<-double()
    start<-0
    Z.mat<-matrix(nrow=n)
    for (ii in 1:n.random){
      r[ii]<-length(Z[[ii]][1,])
      H.mat<-H.mat+obj$sigma*obj$gamma[ii]*Z[[ii]]%*%t(Z[[ii]])
      G.diag[start+1:r[ii]]<-obj$gamma[ii]
      Z.mat<-cbind(Z.mat,Z[[ii]])
      start<-start+r[ii]
    }
    Z.mat<-Z.mat[,-1]
    G.inv<-solve(diag(G.diag))
  }
  base.C<-get.C.base(L=L,X.mat=X,Z=Z.mat,G.inv=G.inv,n=n,q=q,p=p,r=r,n.random=n.random)
  y.star<-t(rmvnorm(B,mean=mean.Xt,cov=H.mat,d=n))
  b.bar.star<-t(apply(X=y.star,MARG=2,FUN=esti.fun,L=L,X.mat=X,Z.mat=Z.mat,
                    C.base=base.C,lambda=obj$sigma/obj$phi,q=q,p=p,r=r,n.random=n.random))
  quants<-matrix(nrow=length(alpha),ncol=q+1)
  for (ii in 1:(q+1))
    quants[,ii]<-quantile(b.bar.star[,1+ii],probs=(1-alpha))

  if (plot.dist){
    par(mfrow=c(2,3))
    hist(b.bar.star[,2],nclass=B/4,prob=TRUE,main="Experimentwise Critical Values",
                                                            xlab="Maximum absolute effect")
    abline(v=quants[,1])
    for (ii in 1:q){
      hist(b.bar.star[,2+ii],nclass=B/4,prob=TRUE,main=paste("Variable",ii),
                                              xlab=paste("Variable",ii," absolute effect"))
```

```
      abline(v=quants[,1+ii])
    }
  }

  if (return.dist)
    res<-list(exp=quants[,1],compar=quants[,1+1:q],dist=b.bar.star)
  else
    res<-list(exp=quants[,1],compar=quants[,1+1:q])
  res

}
```

4.6 Fitting the Alternative LASSO Random Effects Model

See Chapter 10 for details. The function `alt.LASSO()` estimates the dispersion parameters. It calls `diagv()`, `esti.alpha()`, `calc.scores()` and `calc.likelihood()`. The user may want to call individual functions but generally only calls to `alt.LASSO()` will be made.

Utility Function

```
#--------------------------------------------- Form a diagonal square matrix from a vector
diagv<-function(dim,vec=NULL)
#-----------------------------------------------------------------------------------------
#  Forms a dim x dim matrix with vec on the diagonal.
#  If vec==NULL then the identity is returned
{
  mat<-matrix(0,ncol=dim,nrow=dim)
  if (is.null(vec))
    diag(mat)<-1
  else
    diag(mat)<-vec
  mat
}
```

Estimating Random Variance Effects $\boldsymbol{\alpha}$

```
#----------------------------------------------- Estimate random variances effects (alpha)
esti.alpha<-function(y,L,sigma2,two.phi2,tolerance=1e-6,maxit=150,step.size=1,inits=NULL,print.stuff=TRUE)
#-----------------------------------------------------------------------------------------
#Estimates alpha for the alternative LASSO probability model.
#y is the vector of outcomes, L is the design matrix for the LASSO effects, sigma2 is the
#residual variance, two.phi2 is the variance of the LASSO, effects, tolerance is the
#convergnce tolerance for estimates and scores for alpha, maxit is the number of iterations
#allowed, step.size is the length of step to take at each iteration, inits are starting
#values for the alpha, print.stuff is a flag asking if printing of values at each iteration
#and at the end of estimation be performed.

#Function returns a list of 1) the estimated deltas and 2) a list of useful quantities
#-----------------------------------------------------------------------------------------
{
  get.forms<-function(n,q,sigma2,two.phi2,new.alpha,y,L){
  new.delta<-exp(new.alpha)
  D<-diagv(q,new.delta)
  ident.n<-diagv(n)
  ident.q<-diagv(q)
  H<-sigma2*ident.n+L%*%D%*%t(L)
  eig<-eigen(H,symmetric=T)
  eig$values<-ifelse(round(eig$values,10)<=0,0,1/eig$values)
  H.inv<-eig$vectors%*%diag(eig$values)%*%t(eig$vectors)
  w<-t(L)%*%H.inv%*%y
  ww<-w%*%t(w)
  LHL<-t(L)%*%H.inv%*%L
  A.vec<-0.5*diag(LHL)-0.5*diag(ww)+rep(1,q)/two.phi2
  esti.eqn<--new.delta*A.vec+rep(1,q)

  Lambda<-diagv(q,new.delta/two.phi2)+0.5*new.delta%*%t(new.delta)*(LHL*LHL)
  eig<-eigen(Lambda,symmetric=T)
  eig$values<-ifelse(round(eig$values,10)<=0,0,1/eig$values)
  Lambda.inv<-eig$vectors%*%diag(eig$values)%*%t(eig$vectors)

  res<-list(w=w,LHL=LHL,H=H,H.inv=H.inv,Lambda=Lambda,Lambda.inv=Lambda.inv,
                                            esti.eqn=esti.eqn,D=D,d=new.delta,b.vec=A.vec)
  res
  }

  converged<-function(new,prev,tolerance)
    all(abs(new-prev)<tolerance)

  move<-function(prev.alpha,forms,step.size){
    update.d<-prev.alpha+step.size*forms$Lambda.inv%*%forms$esti.eqn
    update.d
  }


#----------------------------------------------------------------------------------------
#                                  The esti.alpha function
#----------------------------------------------------------------------------------------

  n<-dim(L)[1]
  q<-dim(L)[2]
  if (is.null(inits)) new.alpha<-rep(log(two.phi2)-0.5,q) else new.alpha<-inits
  forms<-get.forms(n,q,sigma2,two.phi2,new.alpha,y,L)
  kount<-1
  prev.alpha<-new.alpha
  while ( (kount==1) | (kount<=maxit) &
           !(converged(forms$esti.eqn,rep(0,q),tolerance) & converged(new.alpha,prev.alpha,tolerance)) ){
    prev.alpha<-new.alpha
    if (kount<=5)
      new.alpha<-move(prev.alpha,forms,step.size/50)
  else
    new.alpha<-move(prev.alpha,forms,step.size)
    forms<-get.forms(n,q,sigma2,two.phi2,new.alpha,y,L)
    if (print.stuff)
      print(paste("Iteration:",kount," ","Max difference",round(max(abs(new.alpha-prev.alpha)),5),
                                               "  ","Max derivative:",round(max(abs(forms$esti.eqn)),5)))
    kount<-kount+1
  }

  if (kount==maxit+1){
    print("convergence for alpha failed :-(")
    conv<-FALSE
  }
  else{
    conv<-TRUE
    if (print.stuff)
    print("convergence for alpha succeeded!")
  }

  res<-list(alpha=new.alpha,forms=forms,converged=conv)
  res
}
```

Calculating Approximate Likelihood and Scores

```
#----------------------------------------------Calculate Scores for dispersion parameters
calc.scores<-function(x,alpha=NULL,y,L,q,tolerance=1e-4,maxit=150,type='o')
#----------------------------------------------------------------------------------------
#Calcultes the scores for the dispersion parameters for the alternative LASSO random model
#x = c(ssq, 2\phi^2) and is the dispersion parameter values to use, alpha is the current
#value of the random variance effects (if available), y is the observations, L is the
#design matrix, q is the number of LASSO effects, tolerance is the tolerance to use when
#calculating alpha.hat, maxit is the max number of iterations to use when calculating
```

```
#alpha.hat, type==’o’ signifies to use the observed information for alpha, otherwise the
#expected is used

#Function returns a double (legnth 2) with of scores for x. The double has the attribute
#alpha for the estimated alpha.hat
#-----------------------------------------------------------------------------------------
  {
  get.forms<-function(alpha,forms,a.deriv=NULL,L,y,q,type){
    if (is.null(a.deriv)){
      forms$H2.inv<-forms$H.inv%*%forms$H.inv
      forms$M2<-t(L)%*%forms$H2.inv%*%L
      forms$w2<-t(L)%*%forms$H2.inv%*%y
      forms$MM<-forms$LHL*forms$LHL
      forms$ww<-forms$w%*%t(forms$w)
      forms$wwM<-forms$ww*forms$LHL
      forms$dd<-forms$d%*%t(forms$d)
      forms$ww2<-forms$w%*%t(forms$w2)
      forms$ddM<-forms$dd*forms$LHL
      if (type==’o’){
        forms$Lambda<-diagv(q,forms$d*forms$b.vec) - forms$dd*(0.5*forms$MM-forms$wwM)
        forms$Lambda.inv<-solve(forms$Lambda)
      }
    }
    else{
      temp<-list(sigma2=forms$d*a.deriv$sigma2,two.phi2=forms$d*a.deriv$two.phi2)
      forms$D.star<-list(sigma2=diagv(q,temp$sigma2),two.phi2=diagv(q,temp$two.phi2))
      forms$Delta<-list(sigma2=temp$sigma2%*%t(forms$d),two.phi2=temp$two.phi2%*%t(forms$d))
      forms$MDM<-list(sigma2=forms$LHL%*%forms$D.star$sigma2%*%forms$LHL,
                                   two.phi2=forms$LHL%*%forms$D.star$two.phi2%*%forms$LHL)
      forms$MDw<-list(sigma2=forms$LHL%*%forms$D.star$sigma2%*%forms$w,
                                     two.phi2=forms$LHL%*%forms$D.star$two.phi2%*%forms$w)
    }
    forms
  }

  get.alpha.derivs<-function(x,alpha,forms,q){
    inv.mat<-solve((0.5*forms$MM-forms$wwM)%*%forms$D-solve(forms$D))
    sigma2.deriv<-inv.mat%*%(-0.5*diag(forms$M2)+diag(forms$ww2))
    two.phi2.deriv<- -inv.mat%*%matrix(rep(1,q),ncol=1)/(x[2]^2)
    list(sigma2=sigma2.deriv,two.phi2=two.phi2.deriv)
  }

  get.lambda.derivs.expected<-function(x,alpha,forms,q){
    sigma2.deriv<-forms$D.star$sigma2/x[2]+0.5*(forms$Delta$sigma2+
                     t(forms$Delta$sigma2))*forms$MM-forms$ddM*(forms$M2+forms$MDM$sigma2)
    two.phi2.deriv<- -forms$D/(x[2]^2) + forms$D.star$two.phi2/x[2] +
        0.5*(forms$Delta$two.phi2+t(forms$Delta$two.phi2))*forms$MM-forms$ddM*forms$MDM$two.phi2
    list(sigma2=sigma2.deriv,two.phi2=two.phi2.deriv)
  }

  get.lambda.derivs.observed<-function(x,alpha,forms,q){
    A.mat<- forms$D.star$sigma2*diagv(q,forms$b.vec)+
          diagv(q,forms$d*(-0.5*diag(forms$M2)-
                     0.5*diag(forms$MDM$sigma2)+diag(forms$ww2)+forms$w*forms$MDw$sigma2))
    B.mat<- -(forms$Delta$sigma2+t(forms$Delta$sigma2))*(0.5*forms$MM-forms$wwM)-
          forms$dd*(-forms$LHL*forms$M2+(forms$ww2+t(forms$ww2))*forms$LHL+forms$ww*forms$M2)-
          forms$dd*((forms$ww-forms$LHL)*forms$MDM$sigma2 +
                  (forms$w%*%t(forms$MDw$sigma2)+forms$MDw$sigma2%*%t(forms$w))*forms$LHL)
    sigma2.deriv<-A.mat+B.mat
    A.mat<- forms$D.star$two.phi2*diagv(q,forms$b.vec)+
          diagv(q,forms$d*(-0.5*diag(forms$MDM$two.phi2)+forms$w*forms$MDw$two.phi2-
                                                        matrix(1/(x[2]^2),ncol=1,nrow=q)))
    B.mat<- -(forms$Delta$two.phi2+t(forms$Delta$two.phi2))*(0.5*forms$MM-forms$ww*forms$LHL)-
          forms$dd*((forms$ww-forms$LHL)*forms$MDM$two.phi2 +
              (forms$w%*%t(forms$MDw$two.phi2)+forms$MDw$two.phi2%*%t(forms$w))*forms$LHL)
    two.phi2.deriv<-A.mat+B.mat

    list(sigma2=sigma2.deriv,two.phi2=two.phi2.deriv)
  }
  get.scores<-function(x,alpha,forms,alpha.derivs,Lambda.derivs,q,y){
    tr<-function(mat)
      sum(diag(mat))

    sigma2.score<- -0.5*tr(forms$Lambda.inv%*%Lambda.derivs$sigma2)-0.5*tr(forms$H.inv)-
                    0.5*tr(forms$LHL%*%diagv(q,forms$d*alpha.derivs$sigma2))+
                    0.5*t(y)%*%forms$H2.inv%*%y+
                    0.5*t(forms$w)%*%diagv(q,forms$d*alpha.derivs$sigma2)%*%forms$w-
                    sum(forms$d*alpha.derivs$sigma2)/x[2]+sum(alpha.derivs$sigma2)
    two.phi2.score<- -q/x[2]-0.5*tr(forms$Lambda.inv%*%Lambda.derivs$two.phi2)-
                      0.5*tr(forms$LHL%*%diagv(q,forms$d*alpha.derivs$two.phi2))+
                      0.5*t(forms$w)%*%diagv(q,forms$d*alpha.derivs$two.phi2)%*%forms$w+
                      sum(forms$d)/(x[2]^2)-sum(forms$d*alpha.derivs$two.phi2)/x[2]+
                      sum(alpha.derivs$two.phi2)
    list(sigma2=sigma2.score,two.phi2=two.phi2.score)
  }

  if (is.null(alpha))
    alpha<-esti.alpha(y=y,L=L,sigma2=x[1],two.phi2=x[2],tolerance=tolerance,maxit=maxit,
                                                             inits=NULL,print.stuff=FALSE)
  else
    alpha<-esti.alpha(y=y,L=L,sigma2=x[1],two.phi2=x[2],tolerance=tolerance,maxit=maxit,
                                                            inits=alpha,print.stuff=FALSE)

  forms<-get.forms(alpha=alpha$alpha,forms=alpha$forms,a.deriv=NULL,L=L,y=y,q=q,type=type)
  alpha.derivs<-get.alpha.derivs(x,alpha,forms,q)
  forms<-get.forms(alpha=alpha$alpha,forms=forms,a.deriv=alpha.derivs,L=L,y=y,q=q,type=type)
  if (type=='e')
    Lambda.derivs<-get.lambda.derivs.expected(x,alpha,forms,q)
  else
    Lambda.derivs<-get.lambda.derivs.observed(x,alpha,forms,q)
  scores<- unlist(get.scores(x,alpha$alpha,forms,alpha.derivs,Lambda.derivs,q,y))
  names(scores)<-list("sigma2","two.phi2")
  attr(scores,"alpha")<-alpha

  scores
}

#------------------------------------------------------ Calculates approximate likelihoods
calc.likelihood<-function(x,y,L,alpha=NULL,tolerance=1e-4,maxit,type='o')
#-----------------------------------------------------------------------------------------
#Calculates the approximate likelihood for the alternative LASSO random model.
#x = c(ssq, 2\phi^2) and is the dispersion parameter values to use, y is the outcomes, L is
#the design matrix, alpha is the estimated random variance effects (if available), tolerance
#is the convergence tolerance for the estiamtion of alpha, maxit is the max number of
#iterations to allow in estimating alpha and type='o' indiciates approximation using the
#observed information matrix - otherwise the expected will be used.

#Function returns a double with attributes of the estiamted alpha
#-----------------------------------------------------------------------------------------
{
  get.like<-function(two.phi2,alpha,y){
    Lambda.eig<-eigen(alpha$forms$Lambda,symmetric=TRUE)$values
    Lambda.det<-prod(Lambda.eig[Lambda.eig>0])
    H.eig<-eigen(alpha$forms$H,symmetric=TRUE)$values
    H.det<-prod(H.eig[H.eig>0])
    D.det<-prod(diag(alpha$forms$D))

    like<- -q*log(two.phi2)-0.5*log(Lambda.det)-0.5*log(H.det)-
            0.5*t(y)%*%alpha$forms$H.inv%*%y-
            sum(diag(alpha$forms$D))/two.phi2+sum(alpha$alpha)
    as.double(like)
  }
  if (is.null(alpha))
    alpha<-esti.alpha(y=y,L=L,sigma2=x[1],two.phi2=x[2],tolerance=tolerance,
                                                 maxit=maxit,inits=NULL,print.stuff=FALSE)
  else
    alpha<-esti.alpha(y=y,L=L,sigma2=x[1],two.phi2=x[2],tolerance=tolerance,
                                                maxit=maxit,inits=alpha,print.stuff=FALSE)
  if (type=='o'){
    d<-exp(alpha$alpha)
    alpha$forms$Lambda<-diagv(q,d*alpha$forms$b.vec)-
                                  d%*%t(d)*(0.5*alpha$forms$LHL*alpha$forms$LHL-
                                  alpha$forms$w%*%t(alpha$forms$w)*alpha$forms$LHL)
  }
  like<-get.like(x[2],alpha,y)
  attr(like,"alpha")<-alpha$alpha

  like
}
```

Estimating the Dispersion Parameters

```
#------------------------------------------------------- Estimate the dispersion parameters
alt.LASSO<-function(y,L,init.disp=NULL,tolerance=c(1e-6,1e-3),maxit=c(500,100),
                         step.size=1,type='o',pert=1e-2,print.stuff=TRUE,max.jump=10)
#Estimates the dispersion parameters for the simple LASSO random model.
#y is the outcome vector, L is the design matrix, init.disp is the initial dispersion vector
#c(ssq, 2\phi^2) - if available, tolerance[1] is the tolerance for estimating alpha,
#tolerance[2] is the tolerance for estimating the dispersion, maxit[1] is the max no. of
#iterations for estimating alpha, maxit[2] is the max no. of iterations for estimating
#dispersion, step.size is the size of the step to take when updating dispersions, type=='o'
#signifies estimation using the observed information for alpha - expected info otherwise,
#pert is the value to use when approximating the information for the dispersion, print.stuff
#is a logical (T implies iteration trace is generated) and max.jump is the maximum move a
#dispersion parameter can make.

#Function returns a list: the dispersion estiamtes, the alpha estiamtes, the scores, the
#approx logl and a succesfull convergence flag
#-----------------------------------------------------------------------------------------
{

  calc.hessian<-function(x,alpha,scores,y,L,q,tolerance,maxit,type,pert){
    scores1.1<-scores
    scores2.1<-calc.scores(x=c(x[1]+pert,x[2]),alpha=alpha,y=y,L=L,q=q,
                                                tolerance=tolerance,maxit=maxit,type=type)
    scores1.2<-calc.scores(x=c(x[1],x[2]+pert),alpha=alpha,y=y,L=L,q=q,
                                                tolerance=tolerance,maxit=maxit,type=type)
    hessian<-matrix(,ncol=2,nrow=2)
    hessian[1,1]<-(scores2.1[1]-scores1.1[1])/pert
    hessian[2,2]<-(scores1.2[2]-scores1.1[2])/pert
    hessian[1,2]<-(scores1.2[1]-scores1.1[1])/pert
    hessian[2,1]<-(scores2.1[2]-scores1.1[2])/pert
    hessian[2,1]<-hessian[1,2]<-(hessian[1,2]+hessian[2,1])/2
    if (any(diag(hessian)>=0))
      print(c("Diagonal Hessian positive:",round(diag(hessian),4)))
    hessian
  }

  converged<-function(pt1,pt2,tol)
    all(abs(pt1-pt2)<tol)

  move<-function(disp,scores,hessian,s,max.jump){
    step<- -s*solve(hessian)%*%matrix(scores,ncol=1)
    if (any(abs(step)>max.jump)){
      print("estiamte jumps more than allowed, update reduced")
      step<-ifelse(abs(step)>max.jump,max.jump,step)
    }
    temp<-matrix(disp,ncol=1)+step
    if (any(temp<=0)){
      print("updated dispersion(s) negative: violated estimates reduced to 90%")
      temp[temp<=0,1]<-0.9*disp[temp<=0]
    }
    temp
  }

#-----------------------------------------------------------------------------------------
#                                     alt.LASSO
#-----------------------------------------------------------------------------------------
```

```
  if (is.null(init.disp))
    disp<-c(0.1,0.1)
  else
    disp<-as.double(init.disp)
  names(disp)<-list("sigma2","two.phi2")
  alpha<-esti.alpha(y=y,L=L,sigma2=disp[1],two.phi2=disp[2],tolerance=tolerance[1],
                                              maxit=maxit[1],inits=NULL,print.stuff=FALSE)
  kount<-1
  old.disp<-disp
  scores<-rep(-99,2)
  print(c("iteration",kount-1,"sigma2",round(disp[1],4),"two.phi2",round(disp[2],4)))
  scores<-calc.scores(x=disp,alpha=alpha$alpha,y=y,L=L,q=q,tolerance=tolerance[1],
                                                                 maxit=maxit[1],type=type)
  while ((kount==1) | (kount<=maxit[2]) &
           !(converged(old.disp,disp,tolerance[2])&converged(scores,rep(0,2),tolerance[2]))){

    old.disp<-disp
    alpha<-attr(scores,"alpha")
    hessian<-calc.hessian(x=disp,alpha=alpha$alpha,scores=scores,y=y,L=L,q=q,
                                tolerance=tolerance[1],maxit=maxit[1],type=type,pert=pert)
    disp<-move(disp=disp,scores=scores,hessian=hessian,s=step.size,max.jump)
    scores<-calc.scores(x=disp,alpha=alpha$alpha,y=y,L=L,q=q,
                                          tolerance=tolerance[1],maxit=maxit[1],type=type)
    if (print.stuff)
      print(c("iteration",kount,"sigma2",round(disp[1],4),"two.phi2",round(disp[2],4)))
    kount<-kount+1
  }
  scores<-calc.scores(x=disp,alpha=alpha$alpha,y=y,L=L,q=q,tolerance=tolerance[1],
                                                                 maxit=maxit[1],type=type)
  alpha<-attr(scores,"alpha")$alpha
  if (kount>maxit[2]){
    conv<-as.double(paste(as.double(converged(old.disp,disp,tolerance[2])),
                               as.double(converged(scores,rep(0,2),tolerance[2])),sep=""))
    print(paste("Convergence for Dispersion Parameters Failed with Fault:",conv,sep=" "))
  }
  else{
    if (print.stuff)
      print("Convergence Succeeded")
    conv<-11
  }
  logl<-  calc.likelihood(x=disp,y=y,L=L,alpha=alpha$alpha,tolerance=tolerance[1],
                                                                 maxit=maxit[1],type=type)
  res<-list(dispersion=disp,alpha=alpha,scores=scores,logl=logl,converged=conv)

  res
}
```

4.7 Importance Sampling for the LASSO Random Effects Model

See Chapter 10 for details.

```
#---------------------------------------------- Importance sampling for BP and P(beta>0|y)
LASSO.imp.samps<-function(y,L,disp,B,imp.wts=FALSE)
#-----------------------------------------------------------------------------------------
#Performs importance sampling for posterior expectation (BP) and probabilities
#y is the outcomes, L is the design matrix, B is the number of samples to use, imp.wts is
#boolean - T implies importance weights will be returned.

#Function returns a list with p-values, expected posterior (BP) and the importance weights
#(if asked for)
#-----------------------------------------------------------------------------------------
{
  get.draws<-function(prop.mean,prop.var,B)
    rmvnorm(n=B,mean=prop.mean,cov=prop.var)

  get.joint.dens<-function(samps,y,L,disp,n,q){
    ddexp<-function(X,phi)
```

```
      prod(dexp(abs(X),rate=1/phi)/2)

    my.dmvnorm<-function(X,y,L,sigma2,n)
      prod(dnorm(x=y,mean=L%*%X,sd=sqrt(sigma2)))

    phi<-sqrt(disp[2]/2)
    laps<-apply(samps,MARG=1,FUN=ddexp,phi=phi)
    norms<-apply(samps,MARG=1,FUN=my.dmvnorm,y=as.double(y),L=L,sigma2=disp[1],n=n)
    laps*norms
  }

  get.p.value<-function(samps,wts,q){
    wts.sum<-sum(wts)
    p.vals<-double(length=q)
    for (ii in 1:q){
      indi<-ifelse(samps[,ii]>0,1,0)
      p.vals[ii]<-sum(wts[indi==1])/wts.sum
    }
    p.vals
  }

  get.expect<-function(samps,wts,q){
    temp<-samps*wts
    colSums(temp)/sum(wts)
  }

  calc.BLP<-function(y,L,disp,n,q){
    D.av<-diagv(q,disp[2])
    temp.H.inv<-(diagv(n,1)-L%*%solve(disp[1]*solve(D.av)+t(L)%*%L)%*%t(L))/disp[1]
    BLP<-D.av%*%t(L)%*%temp.H.inv%*%y
    BLP.vari<-D.av-D.av%*%t(L)%*%temp.H.inv%*%L%*%D.av
    t.vals<-BLP/sqrt(diag(BLP.vari))

    list(expect=as.double(BLP),vari=BLP.vari,t.vals=as.double(t.vals))
  }

#------------------------------------------------------------------------------------------
#                              Importance sampling
#------------------------------------------------------------------------------------------
  n<-dim(L)[1]
  q<-dim(L)[2]
  BLP<-calc.BLP(y=y,L=L,disp=disp,n=n,q=q)
  samps<-get.draws(prop.mean=BLP$expect,prop.var=BLP$vari,B=B)
  f.fun<-get.joint.dens(samps=samps,y=y,L=L,disp=disp,n=n,q=q)
  g.fun<-dmvnorm(samps,mean=BLP$expect,cov=BLP$vari)
  wts<-f.fun/g.fun
  p.vals<-get.p.value(samps=samps,wts=wts,q=q)
  expect.wts<-get.expect(samps=samps,wts=wts,q=q)
  if (imp.wts)
    list(p.vals=p.vals,expected=expect.wts,imp.wts=wts)
  else
    list(p.vals=p.vals,expected=expect.wts,imp.wts=NULL)
}
```

References

Afolayan, R. A., Pitchford, W. S., Weatherly, A. W., & Bottema, C. D. K. (2002a). Genetic variation in growth and body dimensions of Jersey and Limousin cross cattle. 1. pre-weaning performance. *Asian-Australian Journal of Animal Sciences* **15**, 1371–1377.

Afolayan, R. A., Pitchford, W. S., Weatherly, A. W., & Bottema, C. D. K. (2002b). Genetic variation in growth and body dimensions of Jersey and Limousin cross cattle. 2. post-weaning dry and wet season performance. *Asian-Australian Journal of Animal Sciences* **15**, 1378–1385.

Andrews, D. F. & Mallows, C. L. (1974). Scale mixtures of normal distributions. *Journal of the Royal Statistical Society Series B* **36**, 99–102.

Apostol, T. M. (1967). *Calculus*, volume 1. John Wiley and Sons, New York, second edition.

Asins, M. J. & Carbonell, E. A. (1988). Detection of linkage between restriction fragment length polymorphism markers and quantitative traits. *Theoretical and Applied Genetics* **76**, 623–626.

A. T. & T. Bell Laboratories (1984). *PORT Mathematical Subroutine Library Manual.*

Ball, R. D. (2001). Bayesian methods for quantitative trait loci mapping based on model selection: Approximate analysis using the Bayesian information criterion. *Genetics* **159**, 1351–1364.

Beckmann, J. S. & Soller, M. (1988). Detection of linkage between marker loci and loci affecting quantitative traits in crosses between segregating populations. *Theoretical and Applied Genetics* **76**, 228–236.

Boer, M. P., ter Braak, C. J. F., & Jansen, R. C. (2002). A penalized likelihood method for mapping epistatic quantitative trait loci with one-dimensional genome searches. *Genetics* **162**, 951–960.

Breiman, L. (1995). Better subset regression using the nonnegative garrote. *Technometrics* **37**, 373–384.

Broman, K. W. (2001). Review of statistical methods for QTL mapping in experimental crosses. *Laboratory Animals* **30**, 44–52.

Broman, K. W. & Speed, T. R. (2002). A model selection approach for the identification of quantitative trait loci in experimental crosses. *Journal of the Royal Statistical Society Series B* **64**, 641–656.

Carbonell, E. A., Gerig, T. M., Balansard, E., & Asins, M. J. (1992). Interval mapping in the analysis of nonadditive quantitative trait loci. *Biometrics* **48**, 305–315.

Chamberlin, A. J., Meuwissen, T. H. E., & Goddard, M. E. (2005). Estimation of the distribution of QTL effects. In *Precedings of the Association for the Advancement of Animal Breeding and Genetics*, volume 16, pages 103–106, Noosa, Queensland, Australia.

Churchill, G. A. & Doerge, R. W. (1994). Empirical threshold values for quantitative trait mapping. *Genetics* **138**, 963–971.

Cook, R. D., Holschuh, N., & Weisberg, S. (1982). A note on an alternative outlier model. *Journal of the Royal Statistical Society Series B* **44**, 370–376.

Cook, R. D. & Weisberg, S. W. (1982). *Residuals and Influence in Regression.* Monographs on Statistics and Applied Probability. Chapman and Hall, New York.

Cullis, B., Smith, A., Verbyla, A., Thompson, R., & Welham, S. (2006). *Mixed Models for Data Analysis.* Draft.

Davison, A. C. & Hinkley, D. V. (1997). *Bootstrap Methods and their Applications.* Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, New York.

de Bruijn, N. G. (1961). *Asymptotic Methods in Analysis.* North-Holland, Amsterdam.

Dodds, K. G., Ball, R., Djorovic, N., & Carson, S. D. (2004). The effect of an imprecise map on interval mapping QTLs. *Genetical Research* **84**, 47–55.

Doerge, R. W. & Churchill, G. A. (1996). Permutation tests for multiple loci affecting a quantitative character. *Genetics* **142**, 285–294.

Doerge, R. W. & Rebaï, A. (1996). Significance thresholds for QTL interval mapping tests. *Heredity* **76**, 459–464.

Draper, N. R., Guttman, I., & Kanemasu, H. (1971). The distribution of certain regression statistics. *Biometrika* **58**, 295–298.

Efron, B., Hastie, T., Johnstone, I., & Tibshirani, R. (2004). Least angle regression. *Annals of Statistics* **32**, 407–451.

Efroymson, M. A. (1960). Multiple regression analysis. In Ralston, A. & Wilf, H. S., editors, *Mathematical methods for digital computers*, volume 1, pages 191–203. Wiley, New York.

Erdélyi, A. (1956). *Asymptotic Expansions.* Dover Publications, New York.

Flury, B. (1997). *A First Course in Multivariate Statistics.* Springer, New York.

Frank, I. E. & Friedman, J. H. (1993). A statistical view of some chemometrics regression tools. *Technometrics* **35**, 109–135.

Fu, W. J. (1998). Penalized regressions: The bridge versus the lasso. *Journal of Computational and Graphical Statistics* **7**, 397–416.

Gassmann, H. I., Deák, I., & Szántai, T. (2002). Computing multivariate normal probabilities: A new look. *Journal of Computational and Graphical Statistics* **11**, 920–949.

Geldermann, H. (1975). Investigations on inheritance of quantitative characters in animals by gene markers I. methods. *Theoretical and Applied Genetics* **46**, 319–330.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, B. R. (1995). *Bayesian Data Analysis.* Texts in Statistical Science. Chapman and Hall/CRC, New York, first edition.

George, A. W., Mengersen, K. L., & Davis, G. P. (2000). Localization of a quantitative trait locus via a Bayesian approach. *Biometrics* **56**, 40–51.

Gianola, D., Perez-Enciso, M., & Toro, M. A. (2003). On marker-assisted prediction of genetic value: Beyond the ridge. *Genetics* **163**, 347–365.

Gilmour, A. R., Thompson, R., & Cullis, B. R. (1995). Average information REML: An efficient algorithm for variance parameter estimation in linear mixed models. *Biometrics* **51**, 1440–1450.

Gogel, B. J., Welham, S. J., Verbyla, A. P., & Cullis, B. R. (2001). Outlier detection in linear mixed effects: Summary of research. Technical Report 2001/P106, BiometricsSA, The University of Adelaide.

Grandvalet, Y. (1998). Least absolute shrinkage is equivalent to quadratic penalization. In Niklasson, L., Bóden, M., & Ziemske, T., editors, *International Conference on Artificial Neural Networks*, volume 1, pages 201–206, Skövde, Sweeden. Springer.

Grandvalet, Y. & Canu, S. (1998). Outcomes of the equivalence of adaptive ridge with least absolute shrinkage. In Kearns, M. S., Solla, S. A., & Cohn, D. A., editors, *Advances in Neural Information Processing Systems*, volume 1, Denver, U.S.A. MIT Press.

Green, P., Falls, K., & Crooks, S. (1990). *Documentation for CRI-MAP, version 2.4.* Washington University School of Medicine, St. Louis, USA.

Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* **82**, 711–732.

Haldane, J. B. S. (1919). The combination of linkage values and the calculation of distance between the loci of linked factors. *Journal of Genetics* **8**, 229–309.

Haley, C. S. & Knott, S. A. (1992). A simple regression method for mapping quantitative trait loci in line crosses using flanking markers. *Heredity* **69**, 315–324.

Haley, C. S., Knott, S. A., & Elsen, J. M. (1994). Mapping quantitative trait loci in crosses between outbred lines using least-squares. *Genetics* **136**, 1195–1207.

Hartl, D. L. & Jones, E. W. (1998). *Genetics: Principles and Analysis.* Jones and Bartlett, London, fourth edition.

Hartwell, L. H., Hood, L., Goldberg, M. L., Reynolds, A. E., Silver, L. M., & Veres, R. C. (2000). Genetics: From genes to genomes. Technical report, McGraw-Hill.

Harville, D. A. (1997). *Matrix Algebra from a Statistician's Perspective.* Springer, New York.

Henderson, C. R. (1950). Estimation of genetic parameters (abstract). *Annals of Mathematical Statistics* **21**, 309–310.

Hoerl, A. E. & Kennard, R. W. (1970a). Ridge regression: Applications to nonorthogonal problems. *Technometrics* **12**, 69–82.

Hoerl, A. E. & Kennard, R. W. (1970b). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* **12**, 55–67.

Hoerl, A. E., Kennard, R. W., & Baldwin, K. F. (1975). Ridge regression: Some simulations. *Communications in Statistics* **4**, 105–123.

Hoeschele, I. & VanRaden, P. M. (1993a). Bayesian-analysis of linkage between genetic-markers and quantitative trait loci .1. prior knowledge. *Theoretical and Applied Genetics*

85, 953–960.

Hoeschele, I. & VanRaden, P. M. (1993b). Bayesian-analysis of linkage between genetic-markers and quantitative trait loci .2. combining prior knowledge with experimental-evidence. *Theoretical and Applied Genetics* **85**, 946–952.

Huang, F. (2003). Prediction error property of the lasso estimator and its generalizations. *Australian and New Zealand Journal of Statistics* **45**, 217–228.

Ihara, N., Takasuga, A., Mizoshita, K., Takeda, H., Sugimoto, M., Mizoguchi, Y., Hirano, T., Itoh, T., Watanabe, T., Reed, K. M., Snelling, W. M., Kappes, S. M., Beattie, C. W., Bennett, G. L., & Sugimoto, Y. (2004). A comprehensive genetic map of the cattle genome based on 3802 microsatellites. *Genome Research* **14**, 1987–1998.

Jansen, R. C. (1992). A general mixture model for mapping quantitative trait loci by using molecular markers. *Theoretical and Applied Genetics* **85**, 252–260.

Jansen, R. C. (1993a). Interval mapping of multiple quantitative trait loci. *Genetics* **135**, 205–211.

Jansen, R. C. (1993b). Maximum-likelihood in a generalized linear finite mixture model by using the EM algorithm. *Biometrics* **49**, 227–231.

Jansen, R. C. (1994). Controlling the type-I and type-II errors in mapping quantitative trait loci. *Genetics* **138**, 871–881.

Jansen, R. C. (1996). A general Monte Carlo method for mapping multiple quantitative trait loci. *Genetics* **142**, 305–311.

Jansen, R. C. & Stam, P. (1994). High-resolution of quantitative traits into multiple loci via interval mapping. *Genetics* **136**, 1447–1455.

Jennrich, R. I. & Sampson, P. F. (1976). Newton-raphson and related algorithms for maximum likelihood variance component estimation. *Technometrics* **18**, 11–17.

Kao, C. H. (2000). On the differences between maximum likelihood and regression interval mapping in the analysis of quantitative trait loci. *Genetics* **156**, 855–865.

Kao, C. H., Zeng, Z. B., & Teasdale, R. D. (1999). Multiple interval mapping for quantitative trait loci. *Genetics* **152**, 1203–1216.

Kass, R. E. & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association* **90**, 773–795.

Kearsey, M. J. & Pooni, H. S. (1996). *The Genetical Analysis of Quantitative Traits*. Chapman and Hall, London.

Kiiveri, H. (2003). A Bayesian approach to variable selection when the number of variables is very large. In Goldstein, D. R., editor, *Science and Statistics: A festscrift for Terry Speed*, volume 40 of *IMS Lecture Notes - Monograph Series*, pages 127–145. Institute of Mathematical Statistics.

Knight, K. & Fu, W. J. (2000). Asymptotics for lasso-type estimators. *Annals of Statistics* **28**, 1356–1378.

Knott, S. A., Elsen, J. M., & Haley, C. S. (1996). Methods for multiple-marker mapping of quantitative trait loci in half-sib populations. *Theoretical and Applied Genetics* **93**, 71–80.

Kosambi, D. D. (1944). The estimation of the map distance from recombination values. *Annals of Eugenics* **12**, 172–175.

Kotz, S., Balakrishnan, N., & Johnson, N. L. (2000). *Continuous Multivariate Distributions Volume 1: Models and Applications*. Wiley Series in Probability and Statistics. Wiley & and Sons, New York, second edition.

Kotz, S., Kozubowski, T. J., & Podgórski (2001). *The Laplace Distribution and Generalizations*. Birkhäuser, Boston.

Lander, E. & Kruglyak, L. (1995). Genetic dissection of complex traits - guidelines for interpreting and reporting linkage results. *Nature Genetics* **11**, 241–247.

Lander, E. S. & Botstein, D. (1989). Mapping mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics* **121**, 185–199.

Lander, E. S. & Green, P. (1987). Construction of multilocus genetic-linkage maps in humans. *Proceedings of the National Academy of Sciences of the United States of America* **84**, 2363–2367.

Lange, K. L., Little, J. A., & Taylor, J. M. G. (1989). Robust statistical modeling using the t distribution. *Journal of the American Statistical Association* **84**, 881–896.

Lindley, D. V. & Smith, A. F. M. (1972). Bayes estimates for the linear model. *Journal of the Royal Statistical Society Series B* **34**, 1–41.

Lokhorst, J. (1999). *The LASSO and Generalised Linear Models*. Honours thesis, The University of Adelaide.

Mardia, K., Kent, J., & Bibby, J. (1979). *Multivariate Analysis*. Academic Press, London.

Martinez, O. & Curnow, R. N. (1992). Estimating the locations and the sizes of the effects of quantitative trait loci using flanking markers. *Theoretical and Applied Genetics* **85**, 480–488.

Martinez, O. & Curnow, R. N. (1994). Missing markers when estimating quantitative trait loci using regression mapping. *Heredity* **73**, 198–206.

McCullagh, P. & Tibshirani, R. (1990). A simple method for the adjustment of profile likelihoods. *Journal of the Royal Statistical Society Series B* **52**, 325–344.

Miller, A. (2002). *Subset Selection in Regression*, volume 95 of *Monographs on Statistics and Applied Probability*. Chapman & Hall/CRC, London, second edition.

Morris, C. A., Cullen, N. G., Pitchford, W. S., Hickey, S. M., Hyndman, D. L., Crawford, A. M., & Bottema, C. D. K. (2003). QTL for birth weight in *bos taurus* cattle. In *Proceedings of the Association for the advancement of animal breeding and genetics*, volume 15, pages 400–403, Melbourne.

Nash, S. G. & Sofer, A. (1996). *Linear and Nonlinear Programming*. Industrial Engineering Series. McGraw-Hill, Singapore.

Nelder, J. A. (1994). The statistics of linear models: back to basics. *Statistics and Computing* **4**, 221–234.

Neter, J., Kutner, M. H., Nachtsheim, C. J., & Wasserman, W. (1996). *Applied Linear*

Statistical Models. Irwin, Chicago, fourth edition.

Öjelund, H., Madsen, H., & Thyregod, P. (2001). Calibration with absolute shrinkage. *Journal of Chemometrics* **15**, 497–509.

Osborne, M. R. (1985). *Finite Algorithms in Optimization and Data Analysis*. Wiley series in probability and mathematical statistics. Wiley, Chichester.

Osborne, M. R., Presnell, B., & Turlach, B. A. (2000). On the LASSO and its dual. *Journal of Computational and Graphical Statistics* **9**, 319–337.

Patterson, H. D. & Thompson, R. (1971). Recovery of interblock information when block sizes are unequal. *Biometrika* **31**, 100–109.

Piepho, H. P. (2001). A quick method for computing approximate thresholds for quantitative trait loci detection. *Genetics* **157**, 425–432.

Piepho, H. P. & Gauch, H. G. (2001). Marker pair selection for mapping quantitative trait loci. *Genetics* **157**, 433–444.

Rebaï, A., Goffinet, B., & Mangin, B. (1994). Approximate thresholds of interval mapping tests for QTL detection. *Genetics* **138**, 235–240.

Rebaï, A., Goffinet, B., & Mangin, B. (1995). Comparing power of different methods for QTL detection. *Biometrics* **51**, 87–99.

Robinson, G. K. (1991). That BLUP is a good thing: The estimation of random effects. *Statistical Science* **6**, 15 –51.

Rosset, S. & Zhu, J. (2004). Corrected proof of the result of 'A prediction error property of the lasso estimator and its generalization' by Huang (2003). *Australian and New Zealand Journal of Statistics* **46**, 505–510.

Satagopan, J. M., Yandell, Y. S., Newton, M. A., & Osborn, T. C. (1996). A Bayesian approach to detect quantitative trait loci using Markov chain Monte Carlo. *Genetics* **144**, 805–816.

Sax, K. (1923). The association of size differences with seed-coat pattern and pigmentation in phaseolus vulgaris. *Genetics* **8**, 552–560.

Searle, S. R., Casella, G., & McCulloch, C. E. (1992). *Variance Components*. Wiley Series in Probability and Statistics. John Wiley & Sons, New York.

Seaton, G., Haley, C. S., Knott, S. A., Kearsey, M., & Visscher, P. M. (2002). QTL Express: mapping quantitative trait loci in of simple and complex pedigrees. *Bioinformatics* **18**, 339–340.

Seber, G. A. F. (1977). *Linear Regression Analysis*. John Wiley and Sons.

Sen, S. & Churchill, G. A. (2001). A statistical framework for quantitative trait mapping. *Genetics* **159**, 371–387.

Sillanpää, M. J. & Arjas, E. (1998). Bayesian mapping of multiple quantitative trait loci from incomplete inbred line cross data. *Genetics* **148**, 1373–1388.

Sillanpää, M. J. & Arjas, E. (1999). Bayesian mapping of multiple quantitative trait loci from incomplete outbred offspring data. *Genetics* **151**, 1605–1619.

Stamey, T. A., Kabalin, J. N., McNeal, J. E., Johnstone, I. M., Freiha, F., Redwine, E. A.,

& Yang, N. (1989). Prostate specific antigen in the diagnosis and treatment of adenocarcinoma of the prostate, II: Radical prostatectomy treated patients. *Journal of Urology* **141**, 1076–1083.

Stoer, J. & Bulirsch, R. (1993). *Introduction to Numerical Analysis*, volume 12 of *Texts in Applied Mathematics*. Springer-Verlag, New York, second edition.

Taylor, J. (2005). *Scale parameter modelling of the t-distribution.* Ph.d, University of Adelaide.

Taylor, J. D. & Verbyla, A. P. (2006). Asymptotic likelihood approximations using a Partial Laplace approximation. *Australian and New Zealand Journal of Statistics* Accepted.

ter Braak, C. J. F., Boer, M. P., & Bink, M. C. A. (2005). Extending Xu's Bayesian model for estimating polygenic effects using markers of the entire genome. *Genetics* **170**, 1435–1438.

Thaller, G. & Hoeschele, I. (1996a). A Monte Carlo method for Bayesian analysis of linkage between single markers and quantitative trait loci .1. methodology. *Theoretical and Applied Genetics* **93**, 1161–1166.

Thaller, G. & Hoeschele, I. (1996b). A Monte Carlo method for Bayesian analysis of linkage between single markers and quantitative trait loci .2. a simulation study. *Theoretical and Applied Genetics* **93**, 1167–1174.

Thompson, R. (1985). A note of restricted maximum likelihood estimation with an alternative outlier model. *Journal of the Royal Statistical Society Series B* **47**, 53–55.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B* **58**, 267–288.

Uimari, P. & Hoeschele, I. (1997). Mapping-linked quantitative trait loci using Bayesian analysis and Markov chain Monte Carlo algorithms. *Genetics* **146**, 735–743.

Uimari, P., Thaller, G., & Hoeschele, I. (1996). The use of multiple markers in a Bayesian method for mapping quantitative trait loci. *Genetics* **143**, 1831–1842.

Venables, W. N. & Ripley, B. D. (2002). *Modern Applied Statistics with S.* Statistics and Computing. Springer, New York, fourth edition.

Verbyla, A. P. (1990). A conditional derivation of residual maximum likelihood. *Australian Journal of Statistics* **32**, 227–230.

Verbyla, A. P., Cullis, B. R., & Thompson, R. (2006). The analysis of quantitative trait loci by simultaneous use of the full linkage map. *In preparation* .

Wang, H., Zhang, Y. M., Li, X. M., Masinde, G. L., Mohan, S., Baylink, D. J., & Xu, S. Z. (2005). Bayesian shrinkage estimation of quantitative trait loci parameters. *Genetics* **170**, 465–480.

Weller, J. I. (1986). Maximum-likelihood techniques for the mapping and analysis of quantitative trait loci with the aid of genetic-markers. *Biometrics* **42**, 627–640.

Whittaker, J. C., Thompson, R., & Denham, M. C. (2000). Marker-assisted selection using ridge regression. *Genetical Research* **75**, 249–252.

Whittaker, J. C., Thompson, R., & Visscher, P. M. (1996). On the mapping of QTL by regression of phenotype on marker-type. *Heredity* **77**, 23–32.

Xu, S. (1995). A comment on the simple regression method for interval mapping. *Genetics* **141**, 1657–1659.

Xu, S. (1998a). Further investigation on the regression method of mapping quantitative trait loci. *Heredity* **80**, 364–373.

Xu, S. (1998b). Iteratively reweighted least squares mapping of quantitative trait loci. *Behavior Genetics* **28**, 341–355.

Xu, S. Z. (2003). Estimating polygenic effects using markers of the entire genome. *Genetics* **163**, 789–801.

Yi, N. J. & Xu, S. Z. (2002). Mapping quantitative trait loci with epistatic effects. *Genetical Research* **79**, 185–198.

Yi, N. J., Xu, S. Z., & Allison, D. B. (2003). Bayesian model choice and search strategies for mapping interacting quantitative trait loci. *Genetics* **165**, 867–883.

Yuan, M. & Lin, Y. (2005). Efficient empirical bayes variable selection and estimation in linear models. *Journal of the American Statistical Association* **100**, 1215–1225.

Zeng, Z. B. (1993). Theoretical basis for separation of multiple linked gene effects in mapping quantitative trait loci. *Proceedings of the National Academy of Sciences of the United States of America* **90**, 10972–10976.

Zeng, Z. B. (1994). Precision mapping of quantitative trait loci. *Genetics* **136**, 1457–1468.

Zeng, Z. B., Kao, C. H., & Basten, C. J. (1999). Estimating the genetic architecture of quantitative traits. *Genetical Research* **74**, 279–289.

Zou, H. & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B* **67**, 301–320.