National Institute for Applied Statistics Research Australia

University of Wollongong, Australia

Working Paper

07-18

A Method of QTL Analysis Utilising Spatial Models for Marker Effects

David Hughes
Supervised by Professor Brian Cullis and Lauren Borg

Copyright © 2017 by the National Institute for Applied Statistics Research Australia, UOW.
Work in progress, no part of this paper may be reproduced without permission from the Institute.

National Institute for Applied Statistics Research Australia, University of Wollongong,
Wollongong NSW 2522, Australia Phone +61 2 4221 5076, Fax +61 2 4221 4998.
Email: karink@uow.edu.au

A method of QTL analysis utilising spatial models for marker effects

David Hughes [4720258]

Supervised by Professor Brian Cullis and Lauren Borg

University of Wollongong

email: dh215@uowmail.edu.au

October 16, 2017

Contents

Abstract

The motivating example of this thesis is an osmotic stress experiment conducted on wheat lines to identify putative quantitative trait loci (QTL) associated with osmotic tolerance. The analytic approach used is an extension of the traditional whole genome average interval mapping (WGAIM) approach (Verbyla et al., 2007), based on markers rather than intervals, which uses the Matrn class of spatial models instead of assuming independence of markers.

In the first chapter, we introduce genetic concepts, such as the structure of DNA, and chromosomes. We establish a definition for recombination, which is a by-product of meiosis, and they key notion that QTL analysis is based on.

Chapter 2 sketches the motivating example, which involves a previous experiment carried out by Dr Rudy Dolferus of the CSIRO. The experiment involved subjecting doubled haploid (DH) lines to an osmotic stress treatment at the sensitive young microspore (YM) stage of development in order to identify quantitative trait loci (QTL) associated with osmotic tolerance. We discuss previous analysis of the experiment, in which Cullis et al. (2015) generate a linkage map from the results.

Chapter 3 delves into a review of previous and current QTL analytical methods. We begin with regression approaches for a single marker model, then move onto the more

complicated flanking marker model. We then move on to the framework for mixture models for QTL analysis, followed by interval and Composite Interval Mapping (CIM), which finally leads into the main technique investigated, WGAIM.

Chapter 4 develops a baseline linear mixed model technique using a spatial marker model for the analysis of QTL. We also build the relevant variance models for the random genetic effects and marker effects.

Finally, in chapter 5, we discuss the estimations and predictions of the model, and establish the marker regression model for computational efficiency.

Acknowledgements

First I'd like to thank Brian Cullis and Lauren Borg for all the help along the course of this project.

A big praise goes to Brian Cullis and Alison Smith for providing a warm environment for me to work on my project in the last few days before the deadline.

I'd also like to thank the Grains Research & Development Corporation (GRDC) for funding me throughout this project with the Undergraduate Honours Scholarship (UHD).

Finally, I'd like to thank Morgan Cooke, who's had to endure a few too many days of the stressful version of me. Without having you by my side, with the love and support you provide, I'd most likely lose it all.

1 Introduction and background genetics

The purpose of this project is to establish a method of Quantitative Trait Loci (QTL) analysis that is an extension of the whole genome average interval mapping (WGAIM) approach described in Verbyla et al. (2007). Unlike WGAIM, the proposed approach avoids the process of forward selection and the resulting selection bias from this process, thus providing more reliable estimates for putative QTL. In addition to this, the WGAIM approach assumes independence of marker effects whereas our proposed approach uses spatial models to model covariance of marker effects within a chromosome.

In the following, we begin with building a foundation of genetic ideologies, providing a broad overview of key relevant concepts of genetics. Next, we discuss the basic framework of QTL analysis, moving from a single gene model to a more complex flanking marker model. We then go on to describe the WGAIM approach to QTL analysis and finally, we discuss our proposal for a more reliable modelling approach.

Due to the advancements of technology, it has become cheaper and faster to analyse genes. There has been an explosion of data relating to DNA sequences, and hence, analysis has shifted from determining what sequences make up DNA, to how these sequences affect the expression of proteins created by certain DNA segments.

QTLs are gene segments found in DNA that correlate with the variation of a phenotype; that is, the observable characteristics of an organism. Examples of phenotypes can range from different hair and eye colours. However, QTLs are specifically linked with traits that vary continuously. An example of such a trait is the yield produced by a specific crop such as wheat.

The determination of these gene segments, or genomic regions, that stimulate the expression of a quantitative trait is imperative in plant and animal breeding. For instance, in plant breeding, identification of putative QTL may lead to being able to plant fewer crops while maximising yield. Furthermore, QTL analysis can also allow for the identification of traits of crops related to disease resistance and tolerance to both physical and biological stresses in the environment.

1.1 Overview of key genetic concepts

First described by Sir Francis Crick (Crick, 1958), the Central Dogma of Biology can be summarised in Figure 1.1, or simply, DNA makes RNA which, in turn, makes proteins.

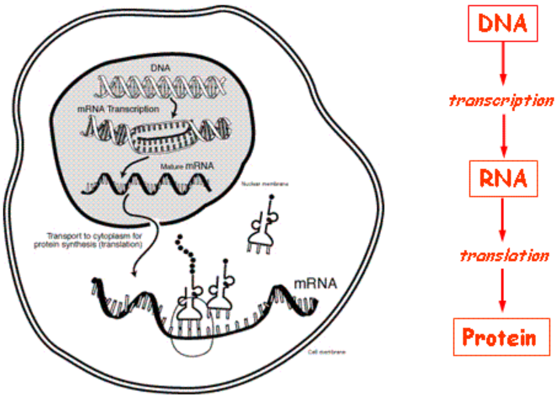


Figure 1.1: (Left) Basic representation of a Eukaryotic Cell which depicts the process of a DNA strand being transcribed into a Messenger RNA (mRNA) strand, which then matures and is transported out of the cell nucleus. When it arrives in the cell cytoplasm, it will undergo translation, and thus a protein molecule specific to the strand of mRNA will be synthesised. (Right) A brief summary of the central dogma of biology portrayed as a flowchart (Berk et al., 1999).

1.1 Overview of key genetic concepts

Deoxyribonucleic acid (DNA) found within an organism is located inside the nucleus of each cell. It is referred to as the "Molecule of Heredity" as it encompasses the genetic instructions which stipulate the biological development of all cellular forms of life. DNA is made of three types of chemical components: a phosphate group, deoxyribose which is a five-carbon-sugar, and four nitrogenous bases, which are either pyrimidines (cytosine, thymine) or purines (adenine, guanine). These two groups of nitrogenous bases differ due to the single-ring structure in pyrimidines compared to the double-ring structure in purines (Figure 1.3). During polymerization, the phosphate group of a subunit (monomer) of DNA fixes onto the deoxyribose of the next group through covalent bonds, creating a long chain (polymer) of information. This polymer is characterised by its double helix shape (Figure 1.2).

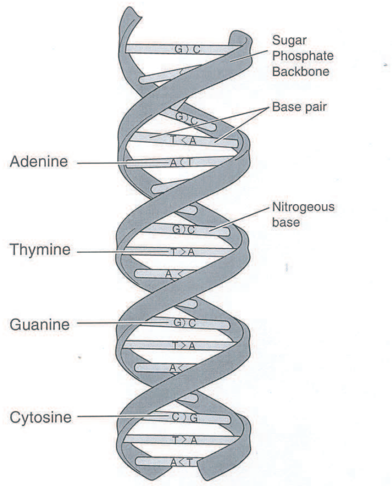


Figure 1.2: Representation of DNA, with its Double-Helix structure with a sugar phosphate backbone. Nitrogenous Bases (Adenine, Thymine, Guanine, and Cytosine) are labelled, with the possible base pairings visible.

Within this double helix structure of DNA, each strand is comprised of a sugar-phosphate backbone which is held in the middle by nitrogenous bases. These nitrogenous bases link up in pairs, with purines bonding with pyrimidines through hydrogen bonding (Figure 1.4). The two strands in this structure are complimentary, that means that at the end of each strand, it is labelled as either 3' or 5' (Figure 1.5), and this will determine the direction that the DNA is read.

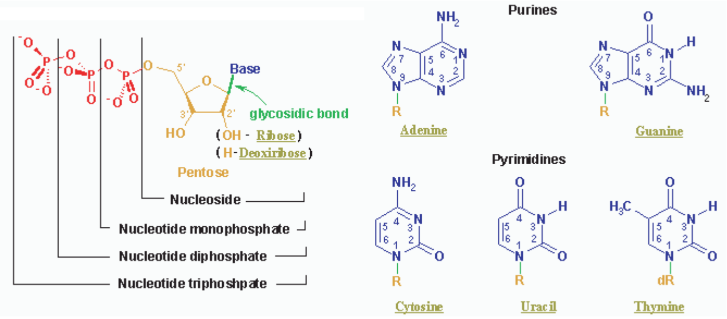


Figure 1.3: (Left) The structure of the sugar phosphate backbone featuring the pentose (sugar) and the phosphate and how these two components bind. (Right) The chemical structures of the nitrogenous bases divided according to Purines (Adenine and Guanine) and Pyrimidines (Cytosine, Thymine and the RNA-found Uracil).

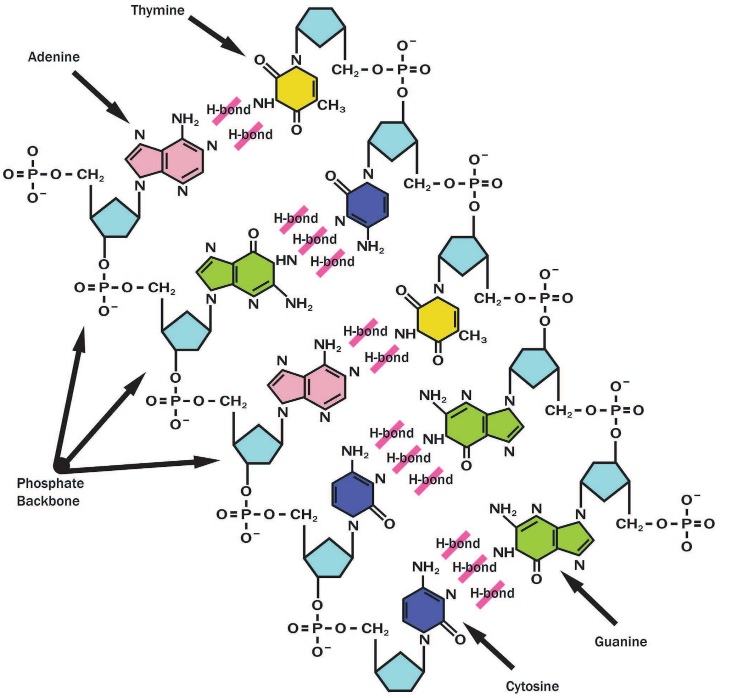


Figure 1.4: The chemical structure of DNA's sugar phosphate backbone, nitrogenous bases, and how the base pairs are formed via Hydrogen Bonds (H-Bonds).

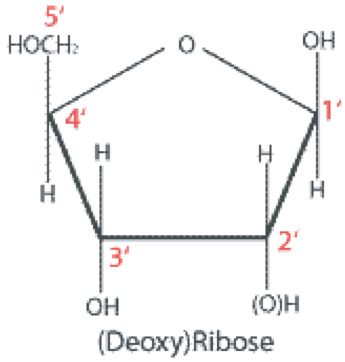


Figure 1.5: The chemical structure of the pentose molecule found in the sugar phosphate backbone of DNA, with the carbon (C-)atoms labelled. A DNA strand is read starting at the 5' C-atom and moves towards the 3' C-atom.

Typically, within the nucleus, DNA is not found as loose strands. In fact, it is packed into thread-like "packages" known as chromosomes (Figure 1.7). It is these packages that are behind the multifaceted organic procedures such as maintenance, development, and reproduction of organisms. Distinct species have differing amounts of chromosomes, and hence, a different genome. For example, human cells are diploid. This means that the nucleus contains two complete sets of chromosomes, one from each parent. Specifically, humans have 23 pairs (46 individual) of chromosomes, which are comprised of 2 sex chromosomes, which determine the gender of the person, and 44 autosomes (i.e. chromosomes that are not sex chromosomes). Interestingly, cattle have 30 pairs of chromosomes (58 autosomes, 2 sex chromosomes). Comparatively, wheat has three genomes, with 7 pairs of chromosomes each, and is known as a hexaploid. However, in usual analysis, it is treated as if it has 21 pairs. This concept of organisms having more than two homologous sets of chromosomes is known as polyploidy (Figure 1.7), and is common in horticultural crops.

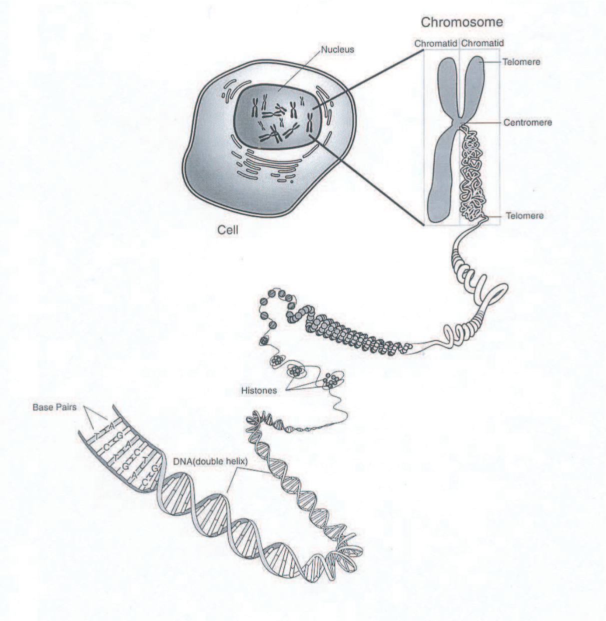


Figure 1.6: A detailed view at the chromosomal hierarchy starting with the highly condensed chromosome found within a cell nucleus. A strand of DNA wraps tightly around the protein histone. This then condenses and forms the resulting chromosome.

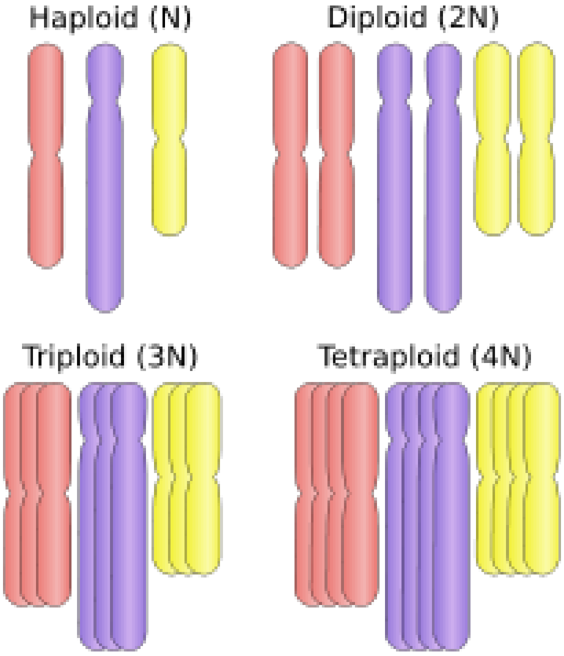


Figure 1.7: Different types of Chromosomal -ploid. Haploid cells carry a single copy of each chromosome, Diploid cells carry two copies, Triploids have three copies, and Tetraploids have four copies.

The central dogma of biology has moulded the foundations of the understandings of biological processes. The cell produces the appropriate protein for the specific process required in several steps. These steps (in order) are transcription, post-transcription, translation, and post-translation. Finally, an active protein is created. This protein may be tasked with triggering a response in a plant in drought situations. Therefore, it is the gene, or more specifically, the segment of DNA sequences that regulates the protein synthesis, and it is this region that is being sought after to be discovered and understood.

1.2 Reproduction

Gene evolution occurs during the reproduction of a species. It is when these changes are introduced that genetic variability is achieved. Breeders can exploit this to improve the genetic composition of an organism, thus introducing important economic attributes to their plants and/or animals. In experimental populations, it is the analysis of new offspring that provides the means for determining genomic regions that affect these attributes. Therefore, it is safe to say that understanding the genetics of reproduction is vital. One form of evolution during reproduction is due to cell division.

1.2.1 Meiosis

There are two forms of cell division: meiosis and mitosis. Mitosis generally occurs when the cell is growing and allows for biological growth and differentiation. The resultant cell from mitosis is two identical daughter cells from a singular parental cell. However, meiosis is somewhat different.

Meiosis is a specialised form of cell division. In a diploid organism (e.g. humans), the resultant cells from this process are four daughter cells, each with half the number of total chromosomes from the progenitor (that is, ancestor) cell. This process involves chromosomal duplication (once) and then cell division (twice).

In diploids, the process begins with cells that have two sets of chromosomes. The DNA duplicates so that each chromosome has identical DNA duplex strands, known as sister chromatids. This pair of chromosomes are also called homologs. From here, the homologs form four duplex strands of DNA with the pairing of non-sister chromatids. After two sequences of cell division, going from diploid (two sets of chromosomes) cells to haploid cells (one set of unpaired chromosomes), the results are four gametes which correspond to each chromosome in the original cell (Figure 1.8).

It is during fertilisation that the diploid characteristic of the organism is restored. In this stage, a male gamete (e.g. sperm) and a female gamete (e.g. ova) fuse together to form the zygote which can develop into an embryo.

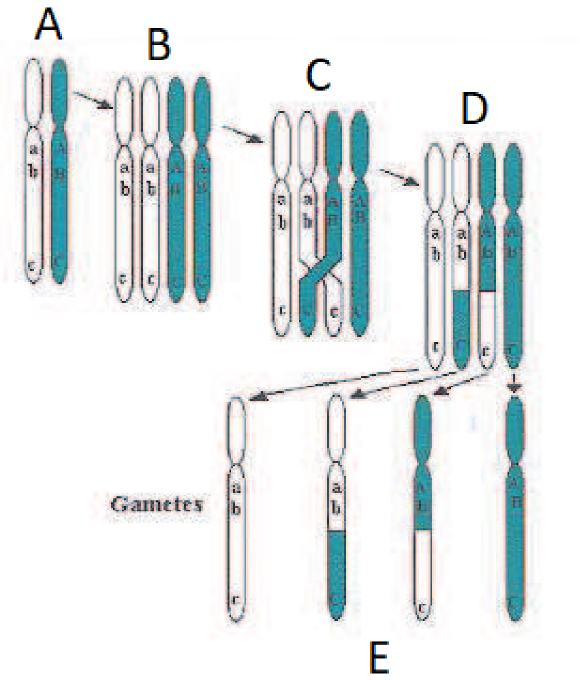


Figure 1.8: The process of meiosis. Starting with A, a parental diploid cell contains two chromosomes; (white) and (blue) corresponding to earlier (grand-)parental cells. In stage B, the chromosomes are duplicated, creating homologs. These non-sister strands are crossed-over in stage C, and are then lined-up and separated in stage D. The final stage E involves cell division forming the four daughter haploid cells (gametes).

1.2.2 Recombination

During cell division, evolutions in the gene segments are introduced during chiasmata (Figure 1.8) in meiosis. This is described as the cross-shaped structure between non-sister chromatids. In the later parts of meiosis, when non-sister chromatids are lined up together, their ends may become transferred. This begins with a process known as crossing-over, and is completed with recombination. These steps encourage the formation of a non-parental haploid when fused in fertilisation, and creates the potential for genetic diversity. Recombination is fundamental in linkage and QTL analysis, as these methods rely on such events. Recombination frequency can be described as the frequency with which a single chromosomal crossover will take place between two genes during meiosis. The greater the frequency between two genetic markers, the further apart they are assumed to be.

1.3 Genetic information

To understand the genes that influence the traits of offspring (progeny), genetic information within the progeny is required. Ideally, the complete DNA sequence for each progeny would be picture-perfect, however this is impossible due to the size and complexity of a DNA sequence, not to mention how sequencing progeny is also an exhaustive task. The solution to this conundrum is to "summarise" information on the genome in some way.

1.3.1 Molecular markers

Molecular markers correspond to genetic differences between individual organisms or species. The marker needs to be polymorphic, that is to say that the marker needs to exhibit variation. Monomorphic markers are contrary to molecular markers, as they do not exhibit variation among offspring. Typically, these markers do not display any information, however in structured populations, they may be informative in some populations, and hence useful.

1.3 Genetic information

In general terms, molecular markers are not genes, but they are genetic “flags” that mark diversity at specific locations on the genome, also known as loci. Markers found close to genes can also be known as gene “tags”.

In genomics, there are several types of markers, specifically categorised as morphological, biochemical, or molecular markers.

Morphological markers are markers that are linked with observed phenotypic traits after the crossing of lines, for example in plant breeding.

Biochemical markers, such as isozymes, are enzymes that have differing amino acid sequences, however they all catalyse the same chemical reaction. These markers are useful for low-level genetic information.

Molecular markers are abundant, and are produced from a broad range of distinct types of DNA mutations. These mutations are specifically substitution mutations (point mutations; transitions, which is replacing a purine for a purine or a pyrimidine for a pyrimidine, or transversions, which is replacing a purine for a pyrimidine, and vice versa), rearrangements, and errors in replication of tandemly repeated DNA.

Generally, there are three forms of molecular markers: Hybridisation-based, Polymerase-Chain-Reaction (PCR)-based, and DNA sequence-based. Hybridisation is the process of merging complementary, single-stranded nucleic acids into a single molecule. However, if there is a slight deviation in sequences between strands, they will not bind. It is based on this that one can determine if an organism has a specific sequence, hence providing researchers with a way to obtain potential markers.

PCR is a tool one utilises to amplify or produce DNA. It begins with denaturing and separating two DNA strands. The enzyme, DNA polymerase, creates copies of the DNA strand by using the original as a template.

1.3 Genetic information

Single Nucleotide Polymorphism (SNP) is a type or DNA sequence-based marker. A SNP is a DNA sequence variation, which occurs when a single nucleotide (purine or pyrimidine) in the genome is altered. SNPs make up 90% of the genetic variations found in humans.

Marker are separated into groups corresponding to chromosomes. These are known as linkage groups.

1.3.2 Checking markers: Segregation analysis

For polymorphic markers, their alleles will be distinct. The relative proportion of each allele, also known as segregation ratios (or odds ratio), will depend on the population type. For example, Doubled Haploid (DH) lines are formed by doubling a haploid cell (i.e. cells with one set of unpaired chromosomes). Therefore, the lines are homozygous at all loci, that is, at particular sections of the gene, each set of duplicated chromosomes will contain the same set of alleles. Furthermore, Recombinant Inbred Lines (RIL) are principally homozygous after sufficient generations, via the process of selfing such as in wheat. For various population types, the odds ratios at a locus that has two possible alleles: A and a are given in Table 1.1.

Table 1.1: Segregation ratios for population types.

Population Type	**Codominant Markers**	**Dominant Markers**
F_2	1:2:1 (aa, Aa, AA)	3:1 (A_1 -, AA)
Backcross	1:1 (Aa, AA)	1:1 (Aa, AA)
RIL or DH	1:1 (aa, AA)	1:1 (aa, AA)

To facilitate the analysis in most mapping populations, genetic markers are given a score which corresponds to the genotype of that marker. Arising from a doubled haploid population, one would arbitrarily record the parental genotypes aa, and AA as 0 and 2, respectively, noting that Aa (= 1), does not occur in this population, as the lines are homozygous. Then, one would subtract this value for the genotype Aa from the other genotypes, giving the maker scores -1 and 1 for aa and AA, respectively, which will be useful in analysis.

2 Motivating example

The data obtained for analysis in this thesis comes from a previous experiment carried out by Dr Rudy Dolferus of the CSIRO, described in Borg et al. (2015). The experiment involved subjecting doubled haploid (DH) lines to an osmotic stress treatment at the sensitive young microspore (YM) stage of development in order to identify quantitative trait loci (QTL) associated with osmotic tolerance.

2.1 Genetic data and map construction

The experiment tested DH lines which result from the crossing of Cranbrook and Halberd wheat varieties, hereafter referred to as the Cranbrook × Halberd (C×H) mapping population. In this population, there were a total 166 DH lines grown in the experiment, as well as the two parents; Cranbrook and Halberd. 165 of these lines were genotyped using a 90K SNP chip.

This chip contained gene-associated SNPs, which provided a sufficiently dense coverage of the wheat genome. These markers were then combined with three phenological markers, that is, known DNA sequences that potentially mark the position of genes associated with phenotypic traits (i.e. QTL), that were available for this DH population. There were a total of 16,231 markers made available for a linkage map construction, the SNP marker names coincide with their index number on the Illumina iSelect 90K SNP array.

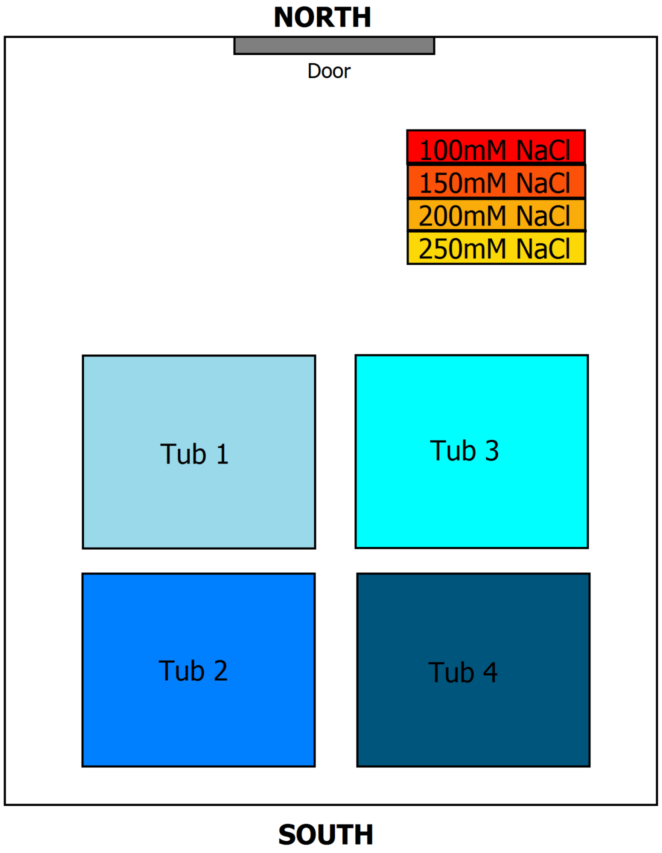


Figure 2.1: Layout of glasshouse from the osmotic stress experiment. The glasshouse contained 4 tubs, in which pots containing lines from the Cranbrook × Halberd DH mapping population were placed. The glasshouse also contained an osmotic treatment vessel, which had varying concentrations of NaCl for the osmotic stress treatment.

Linkage map construction was carried out in the R package, ASMap (Taylor and Butler, 2017). The full details of map construction are given in Cullis et al. (2015), which is briefly described as follows.

Cullis et al. (2015) consider the genotype uniqueness based on the marker data. When genotypes display a high degree of matched pairs of alleles, one can consider the genotypes as being highly related, and hence, enhance segregation distortion. Segregation distortion is when observed genotypic frequencies of a locus do not coincide with the expected Mendelian segregation ratio (Zhan and Xu, 2011). Also, if we were to include these in the analysis of QTL, it may obstruct the identification of putative QTLs in a whole genome approach, due to the base genetic relationship matrix between genotypes not being of full

Table 2.1: Resulting sets of merged clones showing the amount of clones within each set and if clones belong to the same maturity group.

Lines	**Within same maturity group?**
X135 & X151	Yes (Quick)
X145 & X149	No
X148 & X153	Yes (Very Quick)
X17 & X19	Yes (Very Slow)
X44 & X61	No
X4 & X52	Yes (Slow)
X54 & X163	No
X80 & X81	No
X28 & X47 & X64	Yes (Very Quick)
X43 & X55 & X56 & X57	Yes (Quick)

rank (Cullis et al., 2015). These highly related genotypes can be referred to as clones.

Analysis conducted by Cullis et al. (2015) using the `genClones()` function in **ASMap** concluded that there were 10 groups of genotypes that have matched alleles with a frequency of more than 99.5%. These matches were supported by a high number of markers, and as a result, the genotypes within each group will be merged. Table 2.1 has a collection of the resulting sets of merged clones. It was noted that most of the merged sets come from the same maturity group, where lines had been classified into groups previously based on how long their seed took to flower. Interestingly, this result has now added replication into the design, but will now require the construction of additional factors to incorporate the merging of DH lines and also to accommodate DH lines which have phenotypic data but not marker data. This will be done in the following section.

Cullis et al. (2015) then moved onto investigating the quality of the markers based on segregation distortion and the number of missing values. Using the function `profileMark()`, markers that were identified as having a damaging effect on the final quality of the linkage map were removed. This involved removing 33 markers that had a high segregation distortion, 653 markers that were deemed as having too many missing values, and 7,153 markers that were found to be co-located.

Next, the phenological markers had been added back into the map, and formation of the linkage map begins. First, genotypes that were discovered to have cross-overs above a nominal value (e.g. 42) were also dropped from the map object. Six genotypes were discarded.

Before constructing the final linkage map, it was attempted to re-instate some of the markers which were dropped initially due to segregation distortion, missing data, and co-located markers. It was found that 8 from the 33 markers dropped due to segregation distortion, 48 from the 653 markers dropped due to missing information, and 6072 from the 7153 dropped markers due to them being co-located were all re-instated. Finally, after imputing missing values and producing a set of non-redundant markers for each linkage group for QTL analysis, the final linkage map is created denoting the location of the three phenological markers (Figure 2.2).

The final linkage map retains 15,601 markers, reduced to 1,383 non-redundant markers, for the purpose of undertaking the QTL analysis. The number of markers per chromosome ranged from 14 (on Chromosome 3D) to 111 (on Chromosome 2B), with a median number of 74. The overall length of the map was 3,867 cM (Kosambi distance measure, see Appendix for formula), with individual chromosome lengths ranging from 95 to 234. There were 9 genotypes excluded from the QTL analysis, as well as the parents, and a total of 143 genotypes were included with some of these beings groups of genetic clones. See Table 2.2 for a summary of the final linkage map for the C×H mapping population from Cullis et al. (2015).

2.2 Phenotype data set

This experiment involved growing 168 lines (166 DH lines and the two parents; Cranbrook and Halberd) in a glasshouse across two successive runs (See Figure 2.1 for layout). Inside the glasshouse, four hydroponic tubs contained a maximum of 144 pots (arranged in a 16

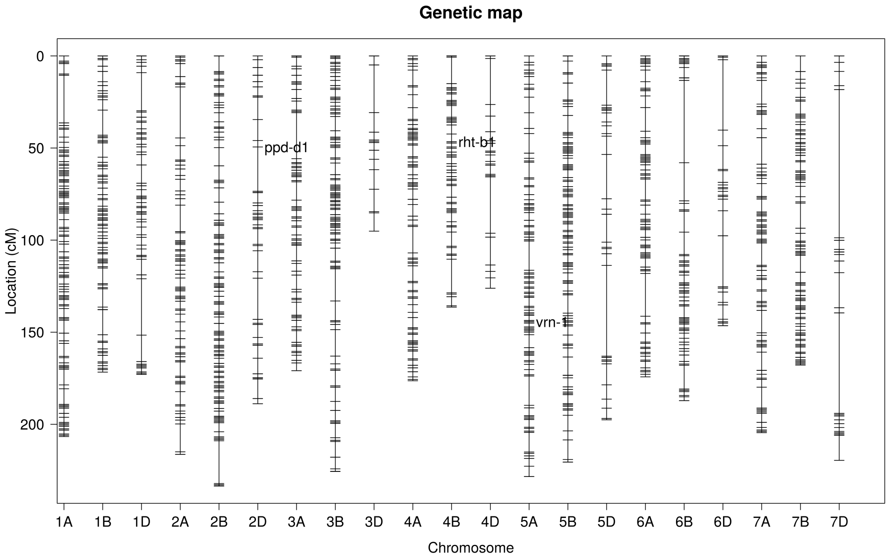


Figure 2.2: Plot of the final map denoting the location of the three phenological markers (Cullis et al., 2015).

Table 2.2: Summary of final linkage map for the C×H mapping population (Cullis et al., 2015), where 'Nmar (full)' is the amount of markers in the final linkage map, and 'Nmar (stats)' is the reduced amount of non-redundant markers that are able to be used for QTL analysis.

	Length (cM)	Nmar (full)	Nmar (stats)
1A	206.7	1313	106
1B	171.7	733	79
1D	172.9	718	53
2A	216.3	1000	74
2B	233.6	1581	111
2D	188.8	450	41
3A	170.9	639	67
3B	225.6	1210	97
3D	95.1	48	14
4A	176.4	941	77
4B	136.3	466	58
4D	126.1	83	22
5A	228.4	579	94
5B	220.5	1544	103
5D	197.6	103	33
6A	174.2	1051	76
6B	187.2	819	60
6D	146.5	119	26
7A	204.5	971	80
7B	167.8	1080	87
7D	219.6	153	25
TOTAL	3867	15601	1383

by 9 rectangular grid).

For the experiment, three seeds per line were absorbed in water for 2 hours, then their surfaces were sterilised for 1 minute using a 0.14% Thiram fungicide solution. Once the seeds were dried off, they were transferred to a wet Whatman 1MM filter paper and then placed in a microwell plate for pre-germination in the dark. Once ready, the seedlings were then moved to pots which contained a fine quartz gravel. Three pots were used for each line, and these pots were then placed into the tubs.

To provide nutrients for the pots, the tubs were subirrigated for 3 minute periods, and then were drained. This process was repeated every 30 minutes, and after 1 week, the water was substituted for a quarter-strength Hoaglands medium. This was elevated to a full strength medium after 2 weeks. The pH of the soil was monitored and maintained within the range of 6.5 - 7.0. Along with the pH, the levels of the nutrient storage tanks was observed weekly, making sure there was always enough to compensate for evaporation. Also, the nutrient solutions were rejuvenated every 2 weeks.

The treatments for the experiment consisted of the factorial combination of the 168 wheat lines with two osmotic stress treatments; namely the presence and absence of osmotic stress, labelled '+' and '-', respectively. The '+' treatment was assigned to two of the three pots per line, and the '-' treatment, or control, to the remaining pot.

During the experiment, lines were tested by subjecting them to an osmotic stress treatment while the seeds were at the YM stage of pollen development. The determination of the young microspore (YM) stage for each seed was based on the auricle distance (AD) measurements. However, this AD at YM varies between each line of the DH population, anywhere from +3 to +8. Although, this variation was balanced out by the knowledge that different florets of a wheat spike, and the spikes on various tillers are asynchronous in flowering.

When the tillers were determined to be at the right stage, the ones that were assigned to a '+' treatment were tagged, and then their pots were placed into the relevant tanks for the application of the treatment.

The treatment involved increasing salt concentration in incremental steps which take the form of 4 separate tanks. These tanks had a Hoagland medium in each, with concentrations of 100, 150, 200, and 250mM of NaCl. During the treatment, the pot would be kept in each of the first 3 tanks for a day, and then the last tank (at the highest concentration) for 2 days. Therefore, the total amount of days of treatment was 5 days. At the end of the 5 days, they were then stepped down to the normal concentration by going backwards over the course of 1 day (one hour per step). The pots were then placed back into tubs until maturity. The spikes from each pot were harvested individually when they reached maturity, and the corresponding spike grain numbers were recorded for each of the treatments (i.e. with osmotic stress ('+') and without ('-')). The spike grain number (SGN) is the trait to be analysed.

2.3 Experimental design

The maturity of each line was assessed from field trials conducted at Yanco and Narrabri experiment locations in New South Wales in previous years, and based on these classifications, lines were then allocated to one of the five maturity blocks (very quick, quick, moderate, slow, and very slow). That is to say that maturity group is aliased with maturity block. The reason behind this allocation was so that one could avoid competition for sunlight between adjacent lines that differed in maturity groups. In Table 2.3, we see how many lines were grown in each maturity block, and in Figures 2.3 and 2.4, we see how these maturity blocks were sorted to tubs inside the glasshouse. The allocation of the maturity blocks to tubs was done in numerical order. However, when maturity blocks were adjacent, a buffer row was included to separate them, again, as to avoid competition

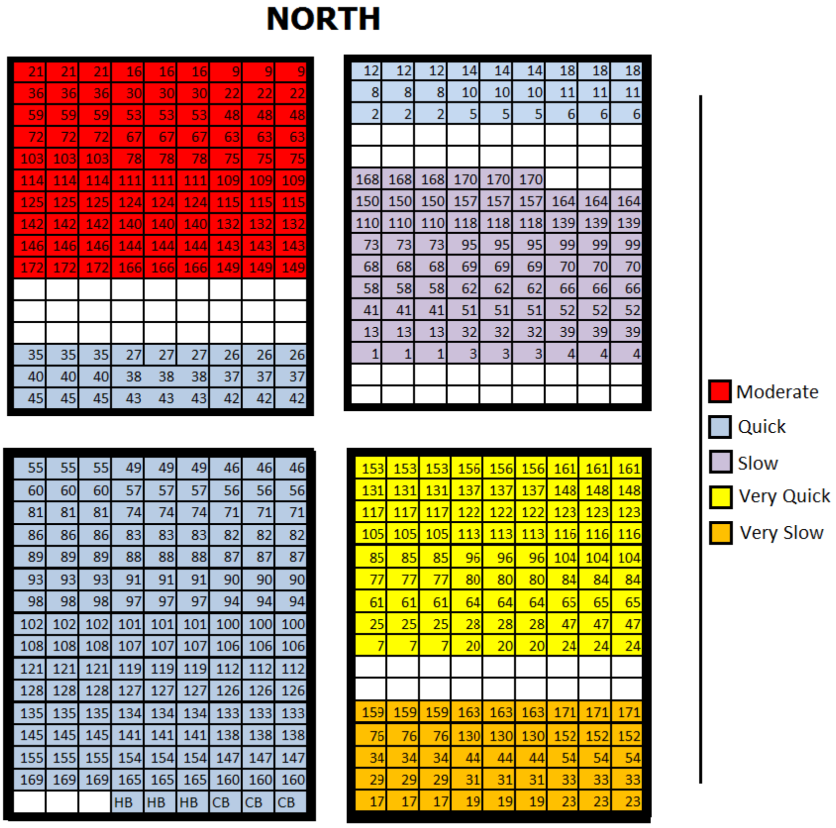
Table 2.3: Number of lines that were originally planted and the amount of lines that have data accompanying them.

	Number of lines grown		**Number of lines with Data**	
Maturity Group	Run 1	Run 2	Run 1	Run 2
Very Quick	27	27	27	27
Quick	66	66	63	64
Moderate	32	32	30	31
Slow	26	26	26	26
Very Slow	17	17	16	17
TOTAL	168	168	162	165

between differing maturity groups.

Each of the two runs of the experiment contained these 5 maturity groups. The DH lines and parents (168 lines) were allocated to a set of 3 row-adjacent pots within their respective maturity block. Within each pot, 3 seeds were planted (total of 9 seeds per line), and when deciding which pot to treat, the two healthiest (i.e best performing) plants remained within the pot, while the other was removed. From here, the osmotic treatment was applied to two randomly chosen pots in each set and the remaining pot was the control. The measurements were taken from the tagged tillers from each plot, however, the number of tillers varied in each pot, as it depended on the number of tillers at the appropriate developmental stage during tagging. It was noted that at the time of measurement, the identity of the pot within each set of 3 pots was not recorded, but the tiller could be identified (in a non-unique way) by a combination of line, treatment, and experiment.

In statistical terms, the observational unit (OU) is the tiller. An OU can be defined by the following factors: Run (which has 2 levels), Matblk (maturity block; 5 levels), Mplot (the set of 3 row-adjacent pots; 48 levels), Splot (pot within each set of row-adjacent pots; 3 levels), and tiller (up to 4 levels). The hydroponic tub associated with the with each OU would also be defined as the factor Tub (4 levels). Finally, there are 2 treatment factors in this experiment. The first being the Treatment ('+' or '-') and the Line (168 levels).



21	21	21	16	16	16	9	9	9
36	36	36	30	30	30	22	22	22
59	59	59	53	53	53	48	48	48
72	72	72	67	67	67	63	63	63
103	103	103	78	78	78	75	75	75
114	114	114	111	111	111	109	109	109
125	125	125	124	124	124	115	115	115
142	142	142	140	140	140	132	132	132
146	146	146	144	144	144	143	143	143
172	172	172	166	166	166	149	149	149
35	35	35	27	27	27	26	26	26
40	40	40	38	38	38	37	37	37
45	45	45	43	43	43	42	42	42

12	12	12	14	14	14	18	18	18
8	8	8	10	10	10	11	11	11
2	2	2	5	5	5	6	6	6
168	168	168	170	170	170			
150	150	150	157	157	157	164	164	164
110	110	110	118	118	118	139	139	139
73	73	73	95	95	95	99	99	99
68	68	68	69	69	69	70	70	70
58	58	58	62	62	62	66	66	66
41	41	41	51	51	51	52	52	52
13	13	13	32	32	32	39	39	39
1	1	1	3	3	3	4	4	4

🟥 Moderate
🟦 Quick
🟪 Slow
🟨 Very Quick
🟧 Very Slow

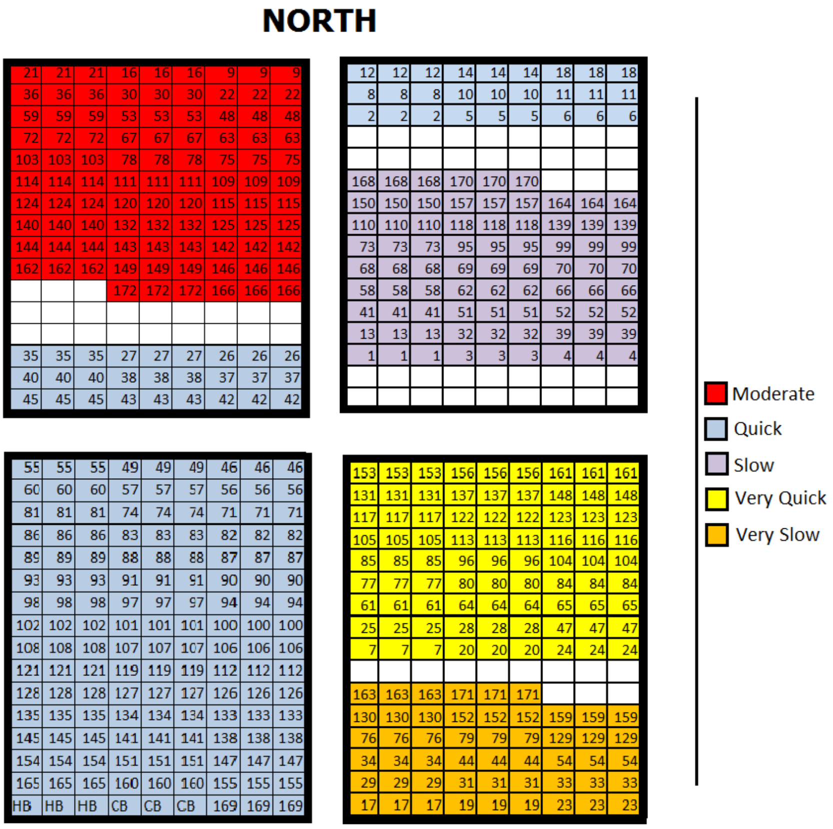
55	55	55	49	49	49	46	46	46
60	60	60	57	57	57	56	56	56
81	81	81	74	74	74	71	71	71
86	86	86	83	83	83	82	82	82
89	89	89	88	88	88	87	87	87
93	93	93	91	91	91	90	90	90
98	98	98	97	97	97	94	94	94
102	102	102	101	101	101	100	100	100
108	108	108	107	107	107	106	106	106
121	121	121	119	119	119	112	112	112
128	128	128	127	127	127	126	126	126
135	135	135	134	134	134	133	133	133
145	145	145	141	141	141	138	138	138
155	155	155	154	154	154	147	147	147
169	169	169	165	165	165	160	160	160
			HB	HB	HB	CB	CB	CB

153	153	153	156	156	156	161	161	161
131	131	131	137	137	137	148	148	148
117	117	117	122	122	122	123	123	123
105	105	105	113	113	113	116	116	116
85	85	85	96	96	96	104	104	104
77	77	77	80	80	80	84	84	84
61	61	61	64	64	64	65	65	65
25	25	25	28	28	28	47	47	47
7	7	7	20	20	20	24	24	24
159	159	159	163	163	163	171	171	171
76	76	76	130	130	130	152	152	152
34	34	34	44	44	44	54	54	54
29	29	29	31	31	31	33	33	33
17	17	17	19	19	19	23	23	23

Figure 2.3: Allocation of lines to pots within maturity blocks for the first run of the osmotic stress experiment. Each individual cell refers to DH lines and parents. The sets of three plots with the same line label are the main-plots.

There were issues regarding the design of the experiment. From Figures 2.3 and 2.4, it was understood that the randomisation of lines to sets of 3 row-adjacent pots was kept the same for each run. On top of this, the maturity blocks also stayed in the same position. Along with these issues, we also do not know the allocation of the treatments to Splots.

It is determined that the design is unreplicated as the experimental units (EU) for lines is the set of 3 row-adjacent pots, and these are all invariant across runs and within maturity blocks. On the other hand, the EU for the treatment by line combination is a pot within the set of 3 row-adjacent pots, but we do not know for certain if the allocation of treatment assigned to pots changed between runs as we cannot identify which pot received which



21	21	21	16	16	16	9	9	9
36	36	36	30	30	30	22	22	22
59	59	59	53	53	53	48	48	48
72	72	72	67	67	67	63	63	63
103	103	103	78	78	78	75	75	75
114	114	114	111	111	111	109	109	109
124	124	124	120	120	120	115	115	115
140	140	140	132	132	132	125	125	125
144	144	144	143	143	143	142	142	142
162	162	162	149	149	149	146	146	146
			172	172	172	166	166	166
35	35	35	27	27	27	26	26	26
40	40	40	38	38	38	37	37	37
45	45	45	43	43	43	42	42	42

12	12	12	14	14	14	18	18	18
8	8	8	10	10	10	11	11	11
2	2	2	5	5	5	6	6	6
168	168	168	170	170	170			
150	150	150	157	157	157	164	164	164
110	110	110	118	118	118	139	139	139
73	73	73	95	95	95	99	99	99
68	68	68	69	69	69	70	70	70
58	58	58	62	62	62	66	66	66
41	41	41	51	51	51	52	52	52
13	13	13	32	32	32	39	39	39
1	1	1	3	3	3	4	4	4

55	55	55	49	49	49	46	46	46
60	60	60	57	57	57	56	56	56
81	81	81	74	74	74	71	71	71
86	86	86	83	83	83	82	82	82
89	89	89	88	88	88	87	87	87
93	93	93	91	91	91	90	90	90
98	98	98	97	97	97	94	94	94
102	102	102	101	101	101	100	100	100
108	108	108	107	107	107	106	106	106
121	121	121	119	119	119	112	112	112
128	128	128	127	127	127	126	126	126
135	135	135	134	134	134	133	133	133
145	145	145	141	141	141	138	138	138
154	154	154	151	151	151	147	147	147
165	165	165	160	160	160	155	155	155
HB	HB	HB	CB	CB	CB	169	169	169

153	153	153	156	156	156	161	161	161
131	131	131	137	137	137	148	148	148
117	117	117	122	122	122	123	123	123
105	105	105	113	113	113	116	116	116
85	85	85	96	96	96	104	104	104
77	77	77	80	80	80	84	84	84
61	61	61	64	64	64	65	65	65
25	25	25	28	28	28	47	47	47
7	7	7	20	20	20	24	24	24
163	163	163	171	171	171			
130	130	130	152	152	152	159	159	159
76	76	76	79	79	79	129	129	129
34	34	34	44	44	44	54	54	54
29	29	29	31	31	31	33	33	33
17	17	17	19	19	19	23	23	23

Figure 2.4: Allocation of lines to pots within maturity blocks for the second run of the osmotic stress experiment. Each individual cell refers to DH lines and parents. The sets of three plots with the same line label are the main-plots.

treatment.

As a result of the linkage map examination in Section 2.1 however, we identified the presence of genetic clones among the lines tested. In this context, a genetic clone is defined as a pair (or in some cases up to 4) DH lines which have identical matching alleles for greater than α% of jointly non-missing markers (we choose $\alpha = 99.5$). As a result, we have minimal partial replication (Cullis et al., 2006) of genotypes for each run.

3 QTL analysis: A review of previous and current approaches

The determination of genomic regions or genes that influence the expression of a quantitative trait is an important endeavour in animal and plant improvement programs. For plants, this can often lead to implementation of efficient selection schemes using approaches generally referred to as marker assisted selection (MAS).

As genome wide dense marker maps are becoming more affordable and available, an important challenge is determining how such information should be incorporated into statistical models for either use in QTL analysis, or more generally, for prediction of additive (or non-additive) genetic values in animal and plant breeding (de los Campos et al., 2009). The latter is referred to as genomic selection (Meuwissen et al., 2001) (GS).

There is a large amount of literature on QTL analysis and more recently the focus has turned from MAS and QTL analysis to GS. The basic idea in MAS is to exploit statistical dependencies (i.e. linkage disequilibrium, LD, that is, the non-random association of alleles at different loci) existing in the joint distribution of markers and QTLs. LD between markers and QTLs has two main objectives, in some way these are not disjoint, not surprisingly, therefore the models used in these two approaches are similar. We refer to those objectives as:

1. QTL analysis in which the aim is to infer genomic regions which affect a quantitative trait, and

2. Genomic selection in which the aim is to predict the genetic merit of individuals

Since the seminal paper of Meuwissen et al. (2001), there has been significant progress made in genomic selection where a realisation was made that unraveling the genetic architecture of a trait via identification of key QTLs is not necessary, nor sufficient for prediction of genetic merit. The concept underlying GS was that a trait (value) is the result of influences of many QTLs with possibly small effects and accurate identification of all of these QTLs would, in general, be very difficult.

Our focus in this project is largely to explore a new approach to QTL analysis, which borrows ideas from the GS literature, and hence, may lead to a more unified and accurate approach. In the next section, we present a review of previous and current approaches to QTL analyses.

3.1 Regression approaches to QTL analysis

Our review will mainly focus on so-called regression approaches to QTL analysis. The most commonly used approaches for QTL analysis in plant breeding applications fall under this broad classification. There are various reasons for their popularity over alternative approaches, such as a mixture modelling approach, or a Bayesian approach such as the one advocated by (Satagopan et al., 1996; Yi and Xu, 2002; Yi, 2004).

Firstly, it has been shown that most regression approaches remain efficient when compared to mixture modelling. Knott (2005) regression-based quantitative trait loci mapping has been shown to be robustly efficient and effective.

Secondly, it is easy to incorporate additional terms which account for non-genetic varia-

Table 3.1: Probabilities for the joint distribution of $(q_i,\ m_i)$, where θ is the recombination fraction between the marker and the QTL.

Marker	QTL	Probability
-1	-1	$1-\theta$
-1	1	θ
1	-1	θ
1	1	$1-\theta$

tion, which is a common occurrence for QTL mapping experiments in plant breeding.

Thirdly, and perhaps most importantly, there are many easy to use, and freely available, QTL analysis packages based on regression methods.

3.1.1 Singer marker regression model

We begin by assuming a very simple genetic model, in which there is a single QTL that affects the trait, $\boldsymbol{y}$, of interest, and further assume we have marker genotype data and trait data on n_g individuals from a single DH population. That is, let m_i be the marker value on individual i, with y_i denoting the trait data for individual i. Further, let q_i be the true QTL genotype for individual i, where q_i takes values -1 and 1 for QTL genotypes qq and QQ, respectively, while m_i takes values -1 and 1 for marker genotypes aa and AA.

Hence, we assume a model for y_i, to be given by

$$y_i|q_i = \mu + \alpha q_i + e_i \tag{3.1}$$

where μ is an overall mean, α is the so-called size of effect of the QTL, e_i is a residual term, with $E(e_i) = 0$.

Since we cannot observe q_i, we consider an approximate model (for y_i) in which we replace q_i by its expectation, $E(q_i|m_i)$. That is, given the conditional probabilities presented in

Table 3.1, we have

$$
\begin{aligned}
E(q_i|m_i = 1) &= -1 \times P(q_i = -1|m_i = 1) + 1 \times P(q_i = 1|m_i = 1) \\
&= -\theta + (1 - \theta) \\
&= (1 - 2\theta) \\
E(q_i|m_i = -1) &= -1 \times P(q_i = -1|m_i = -1) + 1 \times P(q_i = 1|m_i = -1) \\
&= -(1 - \theta) + \theta \\
&= -(1 - 2\theta)
\end{aligned}
$$

and hence,

$$E(q_i|m_i) = (1 - 2\theta)m_i \tag{3.2}$$

Thus, from equations 7.1 and 7.2, we have

$$
\begin{aligned}
E(y_i|m_i) &= \mu + \alpha(1 - 2\theta)m_i + e_i \\
&= \mu + \beta_m m_i + e_i \qquad (3.3)
\end{aligned}
$$

where $\beta_m = \alpha(1 - 2\theta)$ is the regression coefficient of the regression of $\boldsymbol{y}$ on $\boldsymbol{m}$. Hence, it follows that the size and location (through θ) of the QTL are confounded.

We have no way of distinguishing a QTL with a major effect a large distance away from the marker, against a different QTL that has a minor effect a small distance away from a marker. They could both provide similar estimates of β_m. However, this method is obviously easy to implement, and has been often used as a quick scan of the genome to find which markers may be linked to QTLs. It also is worth noting that with increased map density, the disadvantage of this method, regarding confounding, becomes less important.

3.1.2 Flanking marker regression model

The confounding between location and effect of a QTL can be removed if we consider pairs of markers simultaneously. A regression method to do this was proposed independently by Haley and Knott (1992), Martinez and Curnow (1992a). Firstly, we extend our notation used in section 3.1.1, and let m_{ki} be the coded marker genotype of the kth marker for the ith line, $i = 1, ..., n_g$, $k = 1, ..., n_m$. As before, we define q_i to be the QTL genotype for the ith line. We let the marker genotype of the left flanking marker to be m_{Li}, the right flanking marker to be m_{Ri}. We also define θ_{LQ} to be the recombination fraction between the left flanking marker and the QTL, θ_{QR} to be the recombination fraction between the QTL and the right flanking marker, and θ_{LR} to be the recombination fraction between the two flanking markers.

We illustrate this concept in figure 3.1 taken from Whittaker et al. (1996).

Recall the form of the regression model from section 3.1.1 was

$$y_i|m_{ki}, \theta = \mu + \beta_1 E(q_i|m_{ki}, \theta) + e_i \tag{3.4}$$

say, where we noted that

$$E(q_i|m_{ki}, \theta) = (1 - 2\theta)m_{ki}$$

and hence, equation 7.4 becomes

$$y_i|m_{ki}, \theta = \mu + \beta_m m_{ki} + e_i \tag{3.5}$$

say, where $\beta_m = \alpha(1 - 2\theta)$

Here we now replace the expected value of the QTL genotype, given we know the marker genotypes of the two flanking markers. The data is regressed on this expected value and

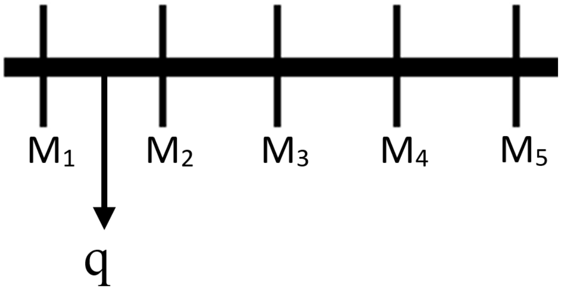


Figure 3.1: Chromosome with markers m_{ki}; $k = 1, ...5$, in map order, q_i between m_{1i} and m_{2i}. The suffix i has been dropped for ease of visualisation. $m_{Li} = m_{1i}$, $m_{Ri} = m_{2i}$.

Table 3.2: Revised expectations for the various combinations of marker genotypes.

m_{Li}	m_{Ri}	$E(q_i\|m_{Li}, m_{Ri}, \theta_{LQ})$
1	1	$\frac{(1-\theta_{LQ}-\theta_{QR})}{(1-\theta_{LR})}$
1	-1	$\frac{(\theta_{QR}-\theta_{LQ})}{\theta_{LR}}$
-1	1	$\frac{-(\theta_{QR}-\theta_{LQ})}{\theta_{LR}}$
-1	-1	$\frac{-(1-\theta_{LQ}-\theta_{QR})}{(1-\theta_{LR})}$

the model is written as

$$y_i|m_{Li}, m_{Ri}, \theta_{LQ} = \mu + \beta_1 E(q_i|m_{Li}, m_{Ri}, \theta_{LQ}) + e_i \tag{3.6}$$

The expectations $E(q_i|m_{Li}, m_{Ri}, \theta_{LQ})$ are presented in Table 3.2.

Using equation 3.6, it is therefore possible to determine if a QTL is present at any location along the genome, not just at the marker loci. These authors used this to scan the genome by fitting the above model at a given set of locations along the genome (i.e. along each chromosome and then for each chromosome) and plotting the residual sum of squares against genetic length to find the most likely position of the QTL. The regression coefficient from the chosen fit, say $\beta_1^{(i)} = a$ represents the size of the QTL effect.

This method was extended by Whittaker et al. (1996) who developed an approach which enabled both the location and size of the QTL to be determined without having to search for possible locations along the genome. This was achieved by simply extending the equation 3.6 to include the two flanking markers. That is, we have

$$y_i|m_{Li}, m_{Ri}, \theta_{LQ} = \mu + \beta_L m_{Li} + \beta_R m_{Ri} + e_i \tag{3.7}$$

and from this we can solve to obtain the size a and location θ_{LQ} of the QTL. To show this result, we let

$$\begin{aligned} \lambda &= \frac{E(q_i|m_{Li}=1, m_{Ri}=1, \theta_{LQ}) + E(q_i|m_{Li}=1, m_{Ri}=-1, \theta_{LQ})}{2} \\ \rho &= \frac{E(q_i|m_{Li}=1, m_{Ri}=1, \theta_{LQ}) + E(q_i|m_{Li}=-1, m_{Ri}=1, \theta_{LQ})}{2} \end{aligned} \tag{3.8}$$

Substitution of the values in table 3.2 into equation 3.8 give

$$\begin{aligned} \lambda &= \frac{\theta_{QR}(1-\theta_{QR})(1-2\theta_{LQ})}{\theta_{LR}(1-\theta_{LR})} \\ \rho &= \frac{\theta_{LQ}(1-\theta_{LQ})(1-2\theta_{QR})}{\theta_{LR}(1-\theta_{LR})} \end{aligned}$$

It can be shown that the four expectations in table 3.2 are given by linear combinations of λ and ρ. In other words

$$E(q_i|m_{Li}, m_{Ri}, \theta_{LQ}) = \lambda m_{Li} + \rho m_{Ri}$$

for all values of m_{Li} and m_{Ri}. Hence, we have

$$\begin{aligned} y_i &= \mu + \beta_1 E(q_i|m_{Li}, m_{Ri}, \theta_{LQ}) + e_i \\ &= \mu + \beta_1(\lambda m_{Li} + \rho m_{Ri}) + e_i \\ &= \mu + \beta_{m_L} m_{Li} + \beta_{m_R} m_{Ri} + e_i \end{aligned}$$

This result shows that the models proposed by Haley and Knott (1992) and Whittaker et al. (1996) are equivalent. The important result is that the parameter θ_{LQ}, which is implicitly a non linear parameter in equation 3.6 has been replaced by an addition parameter and the model is now a linear model in m_{Li} and m_{Ri}.

We can use Trow's formula (Trow, 1913), which is given by

$$\begin{aligned} \theta_{LR} &= \theta_{LQ} + \theta_{QR} - 2\theta_{LQ}\theta_{QR} \\ \Longrightarrow\ 1 - 2\theta_{LR} &= 1 - 2\theta_{LQ} - 2\theta_{QR} + 4\theta_{LQ}\theta_{QR} \\ \Longrightarrow\ 1 - 2\theta_{LR} &= (1 - 2\theta_{LQ})(1 - 2\theta_{QR}) \end{aligned}$$

to eliminate θ_{QR}:

$$\begin{aligned} \beta_{m_L} &= \beta_1\lambda \\ &= \frac{a\theta_{QR}(1-\theta_{QR})(1-2\theta_{LQ})}{\theta_{LR}(1-\theta_{LR})} \\ &= \frac{a(\theta_{LR}-\theta_{LQ})(1-\theta_{LR}-\theta_{LQ})}{\theta_{LR}(1-\theta_{LR})(1-2\theta_{LQ})} \end{aligned}$$

Similarly,

$$\begin{aligned} \beta_{m_R} &= \beta_1\rho \\ &= \frac{a\theta_{LQ}(1-\theta_{LQ})(1-2\theta_{QR})}{\theta_{LR}(1-\theta_{LR})} \\ &= \frac{a\theta_{LQ}(1-\theta_{LQ})(1-2\theta_{LR})}{\theta_{LR}(1-\theta_{LR})(1-2\theta_{LQ})} \end{aligned}$$

Furthermore,

$$\frac{\beta_{m_R}}{\beta_{m_L}} = \frac{\theta_{LQ}(1-\theta_{LQ})(1-2\theta_{LR})}{(\theta_{LR}-\theta_{LQ})(1-\theta_{LR}-\theta_{LQ})}$$

and this gives

$$[\beta_{m_R} + \beta_{m_L}(1-2\theta_{LR})]\theta_{LQ}^2 - [\beta_{m_R} + \beta_{m_L}(1-2\theta_{LR})]\theta_{LQ} + \beta_{m_R}\theta_{LR}(1-\theta_{LR}) = 0 \quad (3.9)$$

Since θ_{LQ} is a recombination fraction, $\theta_{LQ} \in (0, 0.5)$, and so solving equation 3.9 for θ_{LQ} gives the permissible solution

$$\theta_{LQ} = \frac{1}{2}\left(1 - \sqrt{1 - \frac{4\beta_{m_R}\theta_{LR}(1-\theta_{LR})}{\beta_{m_R} + \beta_{m_L}(1-2\theta_{LR})}}\right)$$

We can use this result to show that

$$a^2 = \frac{(\beta_{m_L} + (1-2\theta_{LR})\beta_{m_R})(\beta_{m_R} + (1-2\theta_{LR})\beta_{m_L})}{1-2\theta_{LR}}$$

Thus, the flanking marker regression model can be used to obtain least squares (and near fully efficient) estimation of the size (a) and the location (θ_{LQ}) of the QTL. This obviates the need to undertake a scan between pairs of flanking markers to infer the location (and size) of the QTL. It is important to note that the estimates of β_{m_L} and β_{m_R} must be the same sign, otherwise there is no evidence of a QTL in the given interval.

Another potential issue with this approach, along with the issue of multiple testing and controlling of the family wise error rate; is the occurrence of multiple linked QTLs. That is, if QTLs are present in adjacent intervals, their position and size cannot be determined.

In addition, Xu (1995) showed that the flanking marker or interval mapping regression methods can lead to bias in the estimate of the residual variance or combination of the residual variance as a result of the true QTL genotype being unknown.

Before reviewing other approaches, we note that for most QTL experiments in plants we have statistical (and biological) replication. Lines are usually phenotyped using a designed comparative experiment, which requires an extension of our model to allow for the separate estimation of residual (non-additive) effects from other non-genetic sources of variation.

3.2 Mixture models for QTL analysis

An alternate approach for QTL analysis is based on a maximum likelihood (ML) approach using mixture models. This was the first method proposed for QTL analysis and dates back to Weller (1987). The distribution of y_i, $i = 1..., n_g$, is a mixture of two normal distributions corresponding to each QTL genotype. The likelihood of y_i using the Kth marker, say is given by

$$L = \frac{1}{\left(\sqrt{2\pi\sigma^2}\right)^{n_g}} \prod_{i=1}^{n_g} P(q_i = 1|m_{ki})\text{exp}\left(\frac{-(y_i - \mu_{QQ})^2}{2\sigma^2}\right) + P(q_i = -1|m_{ki})\text{exp}\left(\frac{-(y_i - \mu_{qq})^2}{2\sigma^2}\right) \tag{3.10}$$

where μ_{QQ} and μ_{qq} are the expected values for the two QTL classes; $P(q_i = 1|m_{ki})$ and $P(q_i = -1|m_{ki})$ equal $\frac{1-\theta}{2}$ or $\frac{\theta}{2}$ depending on whether the QTL genotype and the marker genotype are the same or not. We can then maximise L with respect to μ_{QQ}, μ_{qq}, σ^2, and θ. It is possible to test for the presence of a QTL by using a likelihood-ratio test comparing the maximised likelihood with the likelihood obtained when $\theta = 0.5$, that is, when the marker and the (putative) QTL are unlinked.

It is clear that this approach cannot determine the actual location of the QTL in terms of which side of the marker the QTL is on. Thus, Lander and Botstein (1989) extended Weller's approach by using pairs of markers simultaneously. The likelihood is computed for a QTL positioned between two flanking (i.e. adjacent on the map) markers. This likelihood is given by

$$L = \frac{1}{\left(\sqrt{2\pi\sigma^2}\right)^{n_g}} \prod_{i=1}^{n_g} P(q_i = 1|m_{Li}, m_{Ri})\text{exp}\left(\frac{-(y_i - \mu_{QQ})^2}{2\sigma^2}\right) + P(q_i = -1|m_{Li}, m_{Ri})\text{exp}\left(\frac{-(y_i - \mu_{qq})^2}{2\sigma^2}\right) \tag{3.11}$$

Table 6a: Probability of QTL genotype given flanking marker genotype: $P(q_i = z|m_{Li}, m_{Ri})$.

Flanking Marker Genotype		QTL Genotype	
m_{Li}	m_{Ri}	$z = -1$	$z = 1$
-1	-1	$\frac{(1-\theta_{LQ})(1-\theta_{QR})}{(1-\theta_{LR})}$	$\frac{\theta_{LQ}\theta_{QR}}{(1-\theta_{LR})}$
-1	1	$\frac{(1-\theta_{LQ})\theta_{QR}}{\theta_{LR}}$	$\frac{\theta_{LQ}(1-\theta_{QR})}{\theta_{LR}}$
1	-1	$\frac{\theta_{LQ}(1-\theta_{QR})}{\theta_{LR}}$	$\frac{(1-\theta_{LQ})\theta_{QR}}{\theta_{LR}}$
1	1	$\frac{\theta_{LQ}\theta_{QR}}{(1-\theta_{LR})}$	$\frac{(1-\theta_{LQ})(1-\theta_{QR})}{(1-\theta_{LR})}$

where $P(q_i = 1|m_{Li}, m_{Ri})$ is the probability of QTL genotype QQ occurring given the two flanking marker genotypes of m_{Li} and m_{Ri} and are given in table 6a

A likelihood ratio test can be evaluated in a similar way to the single marker model, however the likelihood in 3.11 is difficult to compute. One approach is to compute the likelihood at a number of locations along the chromosome (genome), hence fixing the recombination fractions, making the form of the likelihood more tractable to compute. These likelihoods can then be assessed to determine the most probable location of the QTL. However, as for the flanking marker regression approach, difficulties are encountered when there are linked QTLs and Martinez and Curnow (1992b) showed that this may be a result in a 'ghost' QTL being detected between the two true QTLs.

3.3 Summary of interval and composite interval mapping approaches

Verbyla et al. (2007) note that interval mapping, using either ML or regression methods are the standard approaches for QTL analyses in plants, however they noted that the commonly used software packages which implemented those methods cannot cope with additional sources of variation and this has resulted in the use of so-called two-stage approaches to QTL analysis. This involves the raw data first being subjected to an analysis, to which provides predictions (or means) for each DH line. These means or predictions are then treated as pseudo-data and used in the QTL analyses described in Sections 3.1 and 3.2. This approach is clearly statistically inefficient and piecemeal, notwithstanding the additional issue of multiple testing as the analyses proceed by examining associations at various distances along the linkage map. Furthermore, a one-stage implementation of the genome-scan would become computationally prohibitive, as each requires a re-fit of the base model; each with a new marker or pair of markers (see Boden et al. (2015) for an example).

An extension of interval mapping proposed by Zeng (1994) and Jansen and Stam (1994) is composite interval mapping (CIM). CIM attempts to include a summary of background (additive) genetic variation in the analysis or scan of the genome. The process begins by using (simple) interval mapping to select so-called co-factors. A co-factor, or a set of co-factors are then fitted as additional (fixed) effects in the regression model whilst examining the significance of the 'current' marker or flanking markers. Any co-factor within 10cM of the current marker or marker pairs is not fitted in the model (for the current marker(s)). Thresholds for inclusion of markers in the final model are determined using a parametric bootstrap approach.

3.4 Whole genome approach to QTL analysis

In a major step forward towards resolving some of the issues associated with either simple or CIM, Verbyla et al. (2007) proposed the so-called whole genome average interval mapping (WGAIM) approach. In this approach, all intervals on a linkage map are used simultaneously, thereby avoiding the need for repeated genome-scans. The WGAIM method, however, still uses a forward selection approach, that is, it begins by fitting a simple working model in which it is assumed that a QTL exists in every interval and the QTL sizes are assumed to be random effects, with mean zero, variance σ_α^2, say. A likelihood-ratio test was proposed to determine if a new QTL was to be selected and an alternative outlier detection technique (Thompson, 1985) is used to select the (next) QTL. The QTL is then fitted in the next iteration as a fixed effect. Verbyla et al. (2007) undertook an extensive simulation study which showed the WGAIM approach was more powerful than CIM. There was, however, an issue with the false positive detection rate for WGAIM.

In a subsequent paper, Verbyla et al. (2012) considered an extension of the WGAIM approach which addressed the computational load of the approach when the number of markers was large, and also considered reducing the bias induced as a result of fitting the selected markers (QTLs) as fixed effects. Their extended model also recognised that the WGAIM method could be used for markers rather than intervals when the linkage map had a good coverage and was relatively dense.

4 The analysis of QTL using a spatial marker model

4.1 Introduction

As discussed in 3.4, Verbyla et al. (2012) addressed two issues concerning the WGAIM approach to QTL analysis. Firstly, they improved the computational efficiency of the analysis when the number of markers is large. Secondly, the potential issue of WGAIM was addressed by considering the selected QTL (i.e. markers or pseudo-markers) as random effects. Their approach to reducing the dimensionality was similar to that proposed by VanRaden (2008) for the so-called ridge regression model for use in genome selection. This idea involves using a genomic-relationship matrix, or a variant of this depending on the application. It is therefore clear that although the aims differ between QTL analysis and genomic selection, there is much similarity in the statistical models.

Recently, there has been an increasing interest in the use of so-called spatial models in both QTL analysis and genomic selection. Gianola et al. (2003) considered a range of alternate models for use in marker-assisted selection. These models included effects for chromosomes and correlated or uncorrelated deviation within a chromosome. They used a first order autoregressive process to model the correlation between marker effects on the same chromosome. This model assumed that markers were equidistant (in a genetic sense), although they noted that their model could be extended to use either map or physical distance as the underlying metric for genetic distance. They did not consider the fit of these models to real data.

Yang and Tempelman (2010) considered the class of ante-dependence models (Pourahmadi, 1999) as a model for the covariance of marker effects in the context of genomic selection. Their approach was implemented in a Bayesian framework and they concluded that, on the basis of a simulation study, the models were a "biologically reasonable and computationally tractable method to accommodate linkage disequilibrium (LD)". Further, they stated that the ante-dependence model "should lead to measurably greater gains in accuracy of whole genome selection as greater levels of LD are attained between markers with newly developed SNP marker panels".

In a related approach, there has been interest in the use of spatial models (and related sub-models) for genomic selection. de los Campos et al. (2009) considered the use of a reproducing kernel hilbert space regression (RKHS) model. This function, $g(\cdot)$, is assumed to be within the class of semi-parametric regression models used in the smoothing splines literature and advocated by Green and Silverman (1994). Their choice of penalty function was the gaussian kernel popularised from the RKHS class of models (Wahba, 1990). The gaussian kernel is a one parameter covariance model which assumes an exponential decay as a function of scaled "marker genotype" distance between individuals. They do not provide a formal framework for estimation of the rate constant.

In a similar vein, Ober et al. (2010) use a covariance model which arises from the Matérn class (Stein, 1999) for the genomic relationship between individuals. This model provides more flexibility in capturing the functional dependency of the covariance on the SNP-based genetic distance of individuals. Based on a simulation study, they concluded that there was little difference between the spatial approach and conventional GBLUP.

There is a clear interest in the use of spatial covariance models in genomic selection, but as of yet these models have not been used for the analysis of real data. Spatial models have been proposed as models for the covariance between individuals (based on marker-genotype distances) and the covariance between marker effects within a chromosome. In this chapter, we will therefore develop a general approach to using spatial models in QTL analysis. Our approach is a natural extension of the marker-based approach of WGAIM where we aim to exploit the LD to obtain more accurate estimates of the location of all putative QTLs.

4.2 Statistical models

4.2.1 Baseline linear mixed models

Let $\boldsymbol{y}$ denote the $n \times 1$ vector of (phenotype) data, where n is the number of plots in the experiment (synonymous with trial). We assume that m_d lines were grown in the experiment but marker genotypes (on r markers) are only available for $m < m_d$ lines. For instance, in the motivating example presented in section 2, there were $m_d = 168$ lines but only $m = 143$ lines (including genetic clones) had marker genotypes. This generality allowing $m < m_d$ is important in the context of the analysis of QTL experiments. Many widely used approaches simply discard data from the $(m_d - m)$ lines which do not have marker genotypes (but have phenotypic data). This practice is not only wasteful of information, in terms of the estimation of the sources of non-genetic variation, but can also lead to increased computational burdens if separable spatial models are used for the covariance model of the residuals in QTL experiments conducted in the field (see Cullis and Gleeson (1991) for an example). Hence, in the following we illustrate how it is possible to include data on the $(m_d - m)$ lines in the analysis, but exclude this data from the estimation of the genetic model.

We write the linear mixed model for $\boldsymbol{y}$ as

$$\begin{aligned}\boldsymbol{y} &= \boldsymbol{X}\boldsymbol{\tau} + \boldsymbol{Z}_g\boldsymbol{u}_g + \boldsymbol{Z}_p\boldsymbol{u}_p + \boldsymbol{e} \\ &= \boldsymbol{X}\boldsymbol{\tau} + \boldsymbol{Z}_g\boldsymbol{u}_g + \boldsymbol{e}_p \qquad (4.1)\end{aligned}$$

where $\boldsymbol{e}_p = \boldsymbol{Z}_p\boldsymbol{u}_p + \boldsymbol{e}$, $\boldsymbol{\tau}$ is a vector of fixed effects with associated design matrix $\boldsymbol{X}$; $\boldsymbol{u}_g$ is the $m_d \times 1$ vector of genetic effects with associated design matrix $\boldsymbol{Z}_g$; $\boldsymbol{u}_p$ is a vector of non-genetic or peripheral random effects with associated design matrix $\boldsymbol{Z}_p$ and $\boldsymbol{e}$ is the $n \times 1$ vector of residuals.

We write the fixed effects as $\boldsymbol{\tau} = \left(\boldsymbol{\tau}_0^\top, \boldsymbol{\tau}_g^\top\right)^\top$, where $\boldsymbol{\tau}_g$ is the $(m_d - m) \times 1$ vector of fixed effects corresponding to lines to be excluded from the genetic analysis, and we let $\boldsymbol{X}_g$ denote the associated $n \times (m_d - m)$ design matrix. Thus, $\boldsymbol{X} = [\boldsymbol{X}_0 \ \boldsymbol{X}_g]$, where $\boldsymbol{X}_0$ is the design matrix associated with non-genetic fixed effects $\boldsymbol{\tau}_0$.

In an analogous manner, we consider the partition of $\boldsymbol{u}_g$, which is given by

$$\boldsymbol{u}_g = \begin{bmatrix} \boldsymbol{u}_{g_1} \\ \boldsymbol{u}_{g_2} \end{bmatrix}$$

where $\boldsymbol{u}_{g_1}$ is the $(m_d - m) \times 1$ vector of genetic effects corresponding to the lines to be excluded from the genetic analysis and $\boldsymbol{u}_{g_2}$ is an $m \times 1$ vector of genetic effects to be included in the genetic analysis. The conformal partition of $\boldsymbol{Z}_g$ is given by

$$\boldsymbol{Z}_g = [\boldsymbol{X}_g \ \boldsymbol{Z}_{g2}]$$

Similarly, the genetic variance matrix and its inverse are partitioned conformably as

$$\text{var}\left(\boldsymbol{u}_g\right) = \boldsymbol{G} = \begin{bmatrix} \boldsymbol{G}_{11} & \boldsymbol{G}_{12} \\ \boldsymbol{G}_{21} & \boldsymbol{G}_{22} \end{bmatrix} \quad \text{and} \quad \boldsymbol{G}^{-1} = \begin{bmatrix} \boldsymbol{G}^{11} & \boldsymbol{G}^{12} \\ \boldsymbol{G}^{21} & \boldsymbol{G}^{22} \end{bmatrix} \qquad (4.2)$$

Lastly, we assume that the vectors of random effects and residuals are mutually independent and distributed as multivariate gaussian random variables, with zero means. The variance matrix of $\boldsymbol{u}_p$ is $\boldsymbol{G}_p$ and the variance matrix of $\boldsymbol{e}$ is $\boldsymbol{R}_v$, and hence

$$\begin{aligned} \operatorname{var}(\boldsymbol{e}_p) &= \operatorname{var}(\boldsymbol{Z}_p\boldsymbol{u}_p + \boldsymbol{e}) \\ &= \boldsymbol{Z}_p\boldsymbol{G}_p\boldsymbol{Z}_p^\top + \boldsymbol{R}_v \\ &= \boldsymbol{R}_p \end{aligned}$$

The mixed model equations (MMEs) (Henderson, 1950) for the model in equation 4.1 are given by

$$\begin{bmatrix} \boldsymbol{X}_0^\top\boldsymbol{R}_p^{-1}\boldsymbol{X}_0 & \boldsymbol{X}_0^\top\boldsymbol{R}_p^{-1}\boldsymbol{Z}_{g2} & \boldsymbol{X}_0^\top\boldsymbol{R}_p^{-1}\boldsymbol{X}_g & \boldsymbol{X}_0^\top\boldsymbol{R}_p^{-1}\boldsymbol{X}_g \\ \boldsymbol{Z}_{g2}^\top\boldsymbol{R}_p^{-1}\boldsymbol{X}_0 & \boldsymbol{Z}_{g2}^\top\boldsymbol{R}_p^{-1}\boldsymbol{Z}_{g2} + \boldsymbol{G}^{22} & \boldsymbol{Z}_{g2}^\top\boldsymbol{R}_p^{-1}\boldsymbol{X}_g + \boldsymbol{G}^{21} & \boldsymbol{Z}_{g2}^\top\boldsymbol{R}_p^{-1}\boldsymbol{X}_g \\ \boldsymbol{X}_g^\top\boldsymbol{R}_p^{-1}\boldsymbol{X}_0 & \boldsymbol{X}_g^\top\boldsymbol{R}_p^{-1}\boldsymbol{Z}_{g2} + \boldsymbol{G}^{12} & \boldsymbol{X}_g^\top\boldsymbol{R}_p^{-1}\boldsymbol{X}_g + \boldsymbol{G}^{11} & \boldsymbol{X}_g^\top\boldsymbol{R}_p^{-1}\boldsymbol{X}_g \\ \boldsymbol{X}_g^\top\boldsymbol{R}_p^{-1}\boldsymbol{X}_0 & \boldsymbol{X}_g^\top\boldsymbol{R}_p^{-1}\boldsymbol{Z}_{g2} & \boldsymbol{X}_g^\top\boldsymbol{R}_p^{-1}\boldsymbol{X}_g & \boldsymbol{X}_g^\top\boldsymbol{R}_p^{-1}\boldsymbol{X}_g \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\tau}}_0 \\ \tilde{\boldsymbol{u}}_{g2} \\ \tilde{\boldsymbol{u}}_{g1} \\ \hat{\boldsymbol{\tau}}_g \end{bmatrix} = \begin{bmatrix} \boldsymbol{X}_0^\top\boldsymbol{R}_p^{-1}\boldsymbol{y} \\ \boldsymbol{Z}_{g2}^\top\boldsymbol{R}_p^{-1}\boldsymbol{y} \\ \boldsymbol{X}_g^\top\boldsymbol{R}_p^{-1}\boldsymbol{y} \\ \boldsymbol{X}_g^\top\boldsymbol{R}_p^{-1}\boldsymbol{y} \end{bmatrix} \tag{4.3}$$

Absorbing the equation for $\hat{\boldsymbol{\tau}}_g$ (For reasons due to size, the right hand side of this equation has been summarised after derivation, however see appendix 7.1 which go through a generalised case of matrix absorption including both sides of the equation) the first matrix in 4.3 gives

$$
\begin{aligned}
&\begin{bmatrix}
\boldsymbol{X}_0^\top \boldsymbol{R}_p^{-1} \boldsymbol{X}_0 & \boldsymbol{X}_0^\top \boldsymbol{R}_p^{-1} \boldsymbol{Z}_{g2} & \boldsymbol{X}_0^\top \boldsymbol{R}_p^{-1} \boldsymbol{X}_g \\
\boldsymbol{Z}_{g2}^\top \boldsymbol{R}_p^{-1} \boldsymbol{X}_0 & \boldsymbol{Z}_{g2}^\top \boldsymbol{R}_p^{-1} \boldsymbol{Z}_{g2} + \boldsymbol{G}^{22} & \boldsymbol{Z}_{g2}^\top \boldsymbol{R}_p^{-1} \boldsymbol{X}_g + \boldsymbol{G}^{21} \\
\boldsymbol{X}_g^\top \boldsymbol{R}_p^{-1} \boldsymbol{X}_0 & \boldsymbol{X}_g^\top \boldsymbol{R}_p^{-1} \boldsymbol{Z}_{g2} + \boldsymbol{G}^{12} & \boldsymbol{X}_g^\top \boldsymbol{R}_p^{-1} \boldsymbol{X}_g + \boldsymbol{G}^{11}
\end{bmatrix} \\
&- \begin{bmatrix}
\boldsymbol{X}_0^\top \boldsymbol{R}_p^{-1} \boldsymbol{X}_g \\
\boldsymbol{Z}_{g2}^\top \boldsymbol{R}_p^{-1} \boldsymbol{X}_g \\
\boldsymbol{X}_g^\top \boldsymbol{R}_p^{-1} \boldsymbol{X}_g
\end{bmatrix}
\left(\boldsymbol{X}_g^\top \boldsymbol{R}_p^{-1} \boldsymbol{X}_g\right)^{-1}
\begin{bmatrix}
\boldsymbol{X}_g^\top \boldsymbol{R}_p^{-1} \boldsymbol{X}_0 & \boldsymbol{X}_g^\top \boldsymbol{R}_p^{-1} \boldsymbol{Z}_{g2} & \boldsymbol{X}_g^\top \boldsymbol{R}_p^{-1} \boldsymbol{X}_g
\end{bmatrix} \\
&= \begin{bmatrix}
\boldsymbol{X}_0^\top \boldsymbol{R}_p^{-1} \boldsymbol{X}_0 & \boldsymbol{X}_0^\top \boldsymbol{R}_p^{-1} \boldsymbol{Z}_{g2} & \boldsymbol{X}_0^\top \boldsymbol{R}_p^{-1} \boldsymbol{X}_g \\
\boldsymbol{Z}_{g2}^\top \boldsymbol{R}_p^{-1} \boldsymbol{X}_0 & \boldsymbol{Z}_{g2}^\top \boldsymbol{R}_p^{-1} \boldsymbol{Z}_{g2} + \boldsymbol{G}^{22} & \boldsymbol{Z}_{g2}^\top \boldsymbol{R}_p^{-1} \boldsymbol{X}_g + \boldsymbol{G}^{21} \\
\boldsymbol{X}_g^\top \boldsymbol{R}_p^{-1} \boldsymbol{X}_0 & \boldsymbol{X}_g^\top \boldsymbol{R}_p^{-1} \boldsymbol{Z}_{g2} + \boldsymbol{G}^{12} & \boldsymbol{X}_g^\top \boldsymbol{R}_p^{-1} \boldsymbol{X}_g + \boldsymbol{G}^{11}
\end{bmatrix} \\
&- \begin{bmatrix}
\boldsymbol{X}_0^\top \boldsymbol{R}_p^{-1} \boldsymbol{X}_g \left(\boldsymbol{X}_g^\top \boldsymbol{R}_p^{-1} \boldsymbol{X}_g\right)^{-1} \boldsymbol{X}_g^\top \boldsymbol{R}_p^{-1} \boldsymbol{X}_0 & \boldsymbol{X}_0^\top \boldsymbol{R}_p^{-1} \boldsymbol{X}_g \left(\boldsymbol{X}_g^\top \boldsymbol{R}_p^{-1} \boldsymbol{X}_g\right)^{-1} \boldsymbol{X}_g^\top \boldsymbol{R}_p^{-1} \boldsymbol{Z}_{g2} & \boldsymbol{X}_0^\top \boldsymbol{R}_p^{-1} \boldsymbol{X}_g \left(\boldsymbol{X}_g^\top \boldsymbol{R}_p^{-1} \boldsymbol{X}_g\right)^{-1} \boldsymbol{X}_g^\top \boldsymbol{R}_p^{-1} \boldsymbol{X}_g \\
\boldsymbol{Z}_{g2}^\top \boldsymbol{R}_p^{-1} \boldsymbol{X}_g \left(\boldsymbol{X}_g^\top \boldsymbol{R}_p^{-1} \boldsymbol{X}_g\right)^{-1} \boldsymbol{X}_g^\top \boldsymbol{R}_p^{-1} \boldsymbol{X}_0 & \boldsymbol{Z}_{g2}^\top \boldsymbol{R}_p^{-1} \boldsymbol{X}_g \left(\boldsymbol{X}_g^\top \boldsymbol{R}_p^{-1} \boldsymbol{X}_g\right)^{-1} \boldsymbol{X}_g^\top \boldsymbol{R}_p^{-1} \boldsymbol{Z}_{g2} & \boldsymbol{Z}_{g2}^\top \boldsymbol{R}_p^{-1} \boldsymbol{X}_g \left(\boldsymbol{X}_g^\top \boldsymbol{R}_p^{-1} \boldsymbol{X}_g\right)^{-1} \boldsymbol{X}_g^\top \boldsymbol{R}_p^{-1} \boldsymbol{X}_g \\
\boldsymbol{X}_g^\top \boldsymbol{R}_p^{-1} \boldsymbol{X}_g \left(\boldsymbol{X}_g^\top \boldsymbol{R}_p^{-1} \boldsymbol{X}_g\right)^{-1} \boldsymbol{X}_g^\top \boldsymbol{R}_p^{-1} \boldsymbol{X}_0 & \boldsymbol{X}_g^\top \boldsymbol{R}_p^{-1} \boldsymbol{X}_g \left(\boldsymbol{X}_g^\top \boldsymbol{R}_p^{-1} \boldsymbol{X}_g\right)^{-1} \boldsymbol{X}_g^\top \boldsymbol{R}_p^{-1} \boldsymbol{Z}_{g2} & \boldsymbol{X}_g^\top \boldsymbol{R}_p^{-1} \boldsymbol{X}_g \left(\boldsymbol{X}_g^\top \boldsymbol{R}_p^{-1} \boldsymbol{X}_g\right)^{-1} \boldsymbol{X}_g^\top \boldsymbol{R}_p^{-1} \boldsymbol{X}_g
\end{bmatrix} \\
&= \begin{bmatrix}
\boldsymbol{X}_0^\top \boldsymbol{R}_p^{-1} \boldsymbol{X}_0 & \boldsymbol{X}_0^\top \boldsymbol{R}_p^{-1} \boldsymbol{Z}_{g2} & \boldsymbol{X}_0^\top \boldsymbol{R}_p^{-1} \boldsymbol{X}_g \\
\boldsymbol{Z}_{g2}^\top \boldsymbol{R}_p^{-1} \boldsymbol{X}_0 & \boldsymbol{Z}_{g2}^\top \boldsymbol{R}_p^{-1} \boldsymbol{Z}_{g2} + \boldsymbol{G}^{22} & \boldsymbol{Z}_{g2}^\top \boldsymbol{R}_p^{-1} \boldsymbol{X}_g + \boldsymbol{G}^{21} \\
\boldsymbol{X}_g^\top \boldsymbol{R}_p^{-1} \boldsymbol{X}_0 & \boldsymbol{X}_g^\top \boldsymbol{R}_p^{-1} \boldsymbol{Z}_{g2} + \boldsymbol{G}^{12} & \boldsymbol{X}_g^\top \boldsymbol{R}_p^{-1} \boldsymbol{X}_g + \boldsymbol{G}^{11}
\end{bmatrix} \\
&- \begin{bmatrix}
\boldsymbol{X}_0^\top \boldsymbol{R}_p^{-1} \boldsymbol{X}_g \left(\boldsymbol{X}_g^\top \boldsymbol{R}_p^{-1} \boldsymbol{X}_g\right)^{-1} \boldsymbol{X}_g^\top \boldsymbol{R}_p^{-1} \boldsymbol{X}_0 & \boldsymbol{X}_0^\top \boldsymbol{R}_p^{-1} \boldsymbol{X}_g \left(\boldsymbol{X}_g^\top \boldsymbol{R}_p^{-1} \boldsymbol{X}_g\right)^{-1} \boldsymbol{X}_g^\top \boldsymbol{R}_p^{-1} \boldsymbol{Z}_{g2} & \boldsymbol{X}_0^\top \boldsymbol{R}_p^{-1} \boldsymbol{X}_g \\
\boldsymbol{Z}_{g2}^\top \boldsymbol{R}_p^{-1} \boldsymbol{X}_g \left(\boldsymbol{X}_g^\top \boldsymbol{R}_p^{-1} \boldsymbol{X}_g\right)^{-1} \boldsymbol{X}_g^\top \boldsymbol{R}_p^{-1} \boldsymbol{X}_0 & \boldsymbol{Z}_{g2}^\top \boldsymbol{R}_p^{-1} \boldsymbol{X}_g \left(\boldsymbol{X}_g^\top \boldsymbol{R}_p^{-1} \boldsymbol{X}_g\right)^{-1} \boldsymbol{X}_g^\top \boldsymbol{R}_p^{-1} \boldsymbol{Z}_{g2} & \boldsymbol{Z}_{g2}^\top \boldsymbol{R}_p^{-1} \boldsymbol{X}_g \\
\boldsymbol{X}_g^\top \boldsymbol{R}_p^{-1} \boldsymbol{X}_g \left(\boldsymbol{X}_g^\top \boldsymbol{R}_p^{-1} \boldsymbol{X}_g\right)^{-1} \boldsymbol{X}_g^\top \boldsymbol{R}_p^{-1} \boldsymbol{X}_0 & \boldsymbol{X}_g^\top \boldsymbol{R}_p^{-1} \boldsymbol{X}_g \left(\boldsymbol{X}_g^\top \boldsymbol{R}_p^{-1} \boldsymbol{X}_g\right)^{-1} \boldsymbol{X}_g^\top \boldsymbol{R}^{-1} \boldsymbol{Z}_{g2} & \boldsymbol{X}_g^\top \boldsymbol{R}_p^{-1} \boldsymbol{X}_g
\end{bmatrix} \\
&= \begin{bmatrix}
\boldsymbol{X}_0^\top \boldsymbol{S} \boldsymbol{X}_0 & \boldsymbol{X}_0^\top \boldsymbol{S} \boldsymbol{Z}_{g2} & \boldsymbol{0} \\
\boldsymbol{Z}_{g2}^\top \boldsymbol{S} \boldsymbol{X}_0 & \boldsymbol{Z}_{g2}^\top \boldsymbol{S} \boldsymbol{Z}_{g2} + \boldsymbol{G}^{22} & \boldsymbol{G}^{21} \\
\boldsymbol{0} & \boldsymbol{G}^{12} & \boldsymbol{G}^{11}
\end{bmatrix}
\end{aligned}
$$

where $\boldsymbol{S} = \boldsymbol{R}_p^{-1} - \boldsymbol{R}_p^{-1} \boldsymbol{X}_g \left(\boldsymbol{X}_g^\top \boldsymbol{R}_p^{-1} \boldsymbol{X}_g\right)^{-1} \boldsymbol{X}_g^\top \boldsymbol{R}_p^{-1}$ (assuming $\left(\boldsymbol{X}_g^\top \boldsymbol{R}_p^{-1} \boldsymbol{X}_g\right)$ is full rank), and hence, equation 4.3 becomes, in the usual MME format,

$$\begin{bmatrix} \boldsymbol{X}_0^\top \boldsymbol{S}\boldsymbol{X}_0 & \boldsymbol{X}_0^\top \boldsymbol{S}\boldsymbol{Z}_{g_2} & \boldsymbol{0} \\ \boldsymbol{Z}_{g_2}^\top \boldsymbol{S}\boldsymbol{X}_0 & \boldsymbol{Z}_{g_2}^\top \boldsymbol{S}\boldsymbol{Z}_{g_2} + \boldsymbol{G}^{22} & \boldsymbol{G}^{21} \\ \boldsymbol{0} & \boldsymbol{G}^{12} & \boldsymbol{G}^{11} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\tau}}_0 \\ \tilde{\boldsymbol{u}}_{g_2} \\ \tilde{\boldsymbol{u}}_{g_1} \end{bmatrix} = \begin{bmatrix} \boldsymbol{X}_0^\top \boldsymbol{S}\boldsymbol{y} \\ \boldsymbol{Z}_{g_2}^\top \boldsymbol{S}\boldsymbol{y} \\ \boldsymbol{0} \end{bmatrix} \tag{4.4}$$

Now, as the size of matrices are now somewhat more manageable, we augment our coefficient matrix as follows

$$\begin{bmatrix} \boldsymbol{y}^\top \boldsymbol{S}\boldsymbol{y} & \boldsymbol{y}^\top \boldsymbol{S}\boldsymbol{X}_0 & \boldsymbol{y}^\top \boldsymbol{S}\boldsymbol{Z}_{g_2} & \boldsymbol{0} \\ \boldsymbol{X}_0^\top \boldsymbol{S}\boldsymbol{y} & \boldsymbol{X}_0^\top \boldsymbol{S}\boldsymbol{X}_0 & \boldsymbol{X}_0^\top \boldsymbol{S}\boldsymbol{Z}_{g_2} & \boldsymbol{0} \\ \boldsymbol{Z}_{g_2}^\top \boldsymbol{S}\boldsymbol{y} & \boldsymbol{Z}_{g_2}^\top \boldsymbol{S}\boldsymbol{X}_0 & \boldsymbol{Z}_{g_2}^\top \boldsymbol{S}\boldsymbol{Z}_{g_2} + \boldsymbol{G}^{22} & \boldsymbol{G}^{21} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{G}^{12} & \boldsymbol{G}^{11} \end{bmatrix}$$

we absorb row 4, giving

$$\begin{aligned} &\begin{bmatrix} \boldsymbol{y}^\top \boldsymbol{S}\boldsymbol{y} & \boldsymbol{y}^\top \boldsymbol{S}\boldsymbol{X}_0 & \boldsymbol{y}^\top \boldsymbol{S}\boldsymbol{Z}_{g_2} \\ \boldsymbol{X}_0^\top \boldsymbol{S}\boldsymbol{y} & \boldsymbol{X}_0^\top \boldsymbol{S}\boldsymbol{X}_0 & \boldsymbol{X}_0^\top \boldsymbol{S}\boldsymbol{Z}_{g_2} \\ \boldsymbol{Z}_{g_2}^\top \boldsymbol{S}\boldsymbol{y} & \boldsymbol{Z}_{g_2}^\top \boldsymbol{S}\boldsymbol{X}_0 & \boldsymbol{Z}_{g_2}^\top \boldsymbol{S}\boldsymbol{Z}_{g_2} + \boldsymbol{G}^{22} \end{bmatrix} - \begin{bmatrix} \boldsymbol{0} \\ \boldsymbol{0} \\ \boldsymbol{G}^{21} \end{bmatrix} \left(\boldsymbol{G}^{11}\right)^{-1} \begin{bmatrix} \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{G}^{12} \end{bmatrix} \\ &= \begin{bmatrix} \boldsymbol{y}^\top \boldsymbol{S}\boldsymbol{y} & \boldsymbol{y}^\top \boldsymbol{S}\boldsymbol{X}_0 & \boldsymbol{y}^\top \boldsymbol{S}\boldsymbol{Z}_{g_2} \\ \boldsymbol{X}_0^\top \boldsymbol{S}\boldsymbol{y} & \boldsymbol{X}_0^\top \boldsymbol{S}\boldsymbol{X}_0 & \boldsymbol{X}_0^\top \boldsymbol{S}\boldsymbol{Z}_{g_2} \\ \boldsymbol{Z}_{g_2}^\top \boldsymbol{S}\boldsymbol{y} & \boldsymbol{Z}_{g_2}^\top \boldsymbol{S}\boldsymbol{X}_0 & \boldsymbol{Z}_{g_2}^\top \boldsymbol{S}\boldsymbol{Z}_{g_2} + \boldsymbol{G}^{22} \end{bmatrix} - \begin{bmatrix} \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{G}^{21}(\boldsymbol{G}^{11})^{-1}\boldsymbol{G}^{12} \end{bmatrix} \\ &= \begin{bmatrix} \boldsymbol{y}^\top \boldsymbol{S}\boldsymbol{y} & \boldsymbol{y}^\top \boldsymbol{S}\boldsymbol{X}_0 & \boldsymbol{y}^\top \boldsymbol{S}\boldsymbol{Z}_{g_2} \\ \boldsymbol{X}_0^\top \boldsymbol{S}\boldsymbol{y} & \boldsymbol{X}_0^\top \boldsymbol{S}\boldsymbol{X}_0 & \boldsymbol{X}_0^\top \boldsymbol{S}\boldsymbol{Z}_{g_2} \\ \boldsymbol{Z}_{g_2}^\top \boldsymbol{S}\boldsymbol{y} & \boldsymbol{Z}_{g_2}^\top \boldsymbol{S}\boldsymbol{X}_0 & \boldsymbol{Z}_{g_2}^\top \boldsymbol{S}\boldsymbol{Z}_{g_2} + \left(\boldsymbol{G}^{22} - \boldsymbol{G}^{21}(\boldsymbol{G}^{11})^{-1}\boldsymbol{G}^{12}\right) \end{bmatrix} \end{aligned}$$

Since $\boldsymbol{G}_{22}^{-1} = \boldsymbol{G}^{22} - \boldsymbol{G}^{21}(\boldsymbol{G}^{11})^{-1}\boldsymbol{G}^{12}$ (see Appendix 7.1.2), then the reduced set of MMEs for $\hat{\boldsymbol{\tau}}_0$ and $\tilde{\boldsymbol{u}}_{g2}$ are

$$\begin{bmatrix} \boldsymbol{X}_0^\top \boldsymbol{S}\boldsymbol{X}_0 & \boldsymbol{X}_0^\top \boldsymbol{S}\boldsymbol{Z}_{g2} \\ \boldsymbol{Z}_{g2}^\top \boldsymbol{S}\boldsymbol{X}_0 & \boldsymbol{Z}_{g2}^\top \boldsymbol{S}\boldsymbol{Z}_{g2} + \boldsymbol{G}_{22}^{-1} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\tau}}_0 \\ \tilde{\boldsymbol{u}}_{g2} \end{bmatrix} = \begin{bmatrix} \boldsymbol{X}_0^\top \boldsymbol{S}\boldsymbol{y} \\ \boldsymbol{Z}_{g2}^\top \boldsymbol{S}\boldsymbol{y} \end{bmatrix} \tag{4.5}$$

The MMES in equation 4.5 are the same as those for the linear mixed model given by

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\tau} + \boldsymbol{Z}_{g_2}\boldsymbol{u}_{g_2} + \boldsymbol{e}_p \tag{4.6}$$

Therefore, instead of using the linear mixed model in equation 4.1, in which the vector of random genetic effects is of length m_d and corresponds to all lines grown in the trial, we could fit the linear mixed model commensurate with the MMEs in 4.5, which is 4.6. In this model, the vector of random genetic effects, $\boldsymbol{u}_{g_2}$, is of length m and corresponds to only those lines of interest, that is, those to be used in the genetic (i.e. QTL) analysis which have marker genotypes. Additionally, the model includes fixed effects, $\boldsymbol{\tau}_g$, corresponding to the lines to be excluded. □

4.2.2 Linear mixed model including marker genotypes

Given the results in section 4.2.1, we can therefore write the baseline model for $\boldsymbol{y}$ as

$$\begin{aligned} \boldsymbol{y} &= \boldsymbol{X}\boldsymbol{\tau} + \boldsymbol{Z}_g\boldsymbol{u}_g + \boldsymbol{Z}_p\boldsymbol{u}_p + \boldsymbol{e} \\ &= \boldsymbol{X}\boldsymbol{\tau} + \boldsymbol{Z}_g\boldsymbol{u}_g + \boldsymbol{e}_p \end{aligned} \tag{4.7}$$

where $\boldsymbol{\tau}$ is a vector of fixed effects with associated design matrix $\boldsymbol{X}$; $\boldsymbol{u}_g$ is the $m \times 1$ vector of random genetic effects corresponding to those lines with marker genotypes and has an associated $n \times m$ design matrix $\boldsymbol{Z}_g$; $\boldsymbol{u}_p$ is a vector of non-genetic, or peripheral, random effects with associated design matrix $\boldsymbol{Z}_p$, and $\boldsymbol{e}$ is the $n \times 1$ vector of residuals.

The fixed effects are partitioned at $\boldsymbol{\tau}$,

$$\boldsymbol{\tau} = \begin{bmatrix} \boldsymbol{\tau}_0 \\ \boldsymbol{\tau}_g \end{bmatrix}$$

where $\boldsymbol{\tau}_g$ is the $(m_d - m) \times 1$ vector of fixed effects corresponding to the lines without marker data, and we let $\boldsymbol{X}_g$ denote the associated $n \times (m_d - m)$ design matrix. Thus, $\boldsymbol{X} = \begin{bmatrix} \boldsymbol{X}_0 & \boldsymbol{X}_g \end{bmatrix}$, where $\boldsymbol{X}_0$ is the design matrix associated with the non-genetic fixed effects $\boldsymbol{\tau}_0$.

Further, we assume that

$$\begin{bmatrix} \boldsymbol{u}_p \\ \boldsymbol{e} \end{bmatrix} \sim N \left(\begin{bmatrix} \boldsymbol{0} \\ \boldsymbol{0} \end{bmatrix}, \begin{bmatrix} \boldsymbol{G}_p & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{R}_v \end{bmatrix} \right)$$

Thus, if $\boldsymbol{e}_p = \boldsymbol{Z}_p \boldsymbol{u}_p + \boldsymbol{e}$ as before, then

$$\text{var}(\boldsymbol{e}_p) = \boldsymbol{R}_p = \boldsymbol{Z}_p \boldsymbol{G}_p \boldsymbol{Z}_p^\top + \boldsymbol{R}_v$$

To allow for the inclusion of marker genotypes, we consider the model for $\boldsymbol{u}_g$ given by

$$\boldsymbol{u}_g = \boldsymbol{u}_a + \boldsymbol{u}_e \tag{4.8}$$

where the two terms represent the additive and non-additive (or residual) genetic effects, respectively. Then we propose that the additive genetic effects be modelled as a linear function of coded marker genotypes. Hence, we have

$$\boldsymbol{u}_a = \boldsymbol{M}\boldsymbol{\alpha} + \boldsymbol{u}_\epsilon \tag{4.9}$$

where $\boldsymbol{M}$ is the $m \times r$ matrix of marker covariate data, where the columns of $\boldsymbol{M}$ are in map order. We note $\boldsymbol{M} = \begin{bmatrix} \boldsymbol{M}_1 & \boldsymbol{M}_2 & \dots & \boldsymbol{M}_c \end{bmatrix}$, where each $\boldsymbol{M}_i$ is a $m \times r_i$ matrix corresponding to LG i. The vector $\boldsymbol{\alpha}$ is the $r \times 1$ vector of random marker effects (regression coefficients) and is conformably partitioned (i.e. $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1^\top, \dots, \boldsymbol{\alpha}_c^\top)^\top$), and $\boldsymbol{u}_\epsilon$ is the $m \times 1$ vector of lack of fit effects for the marker regression. For simple mapping populations, such as DH populations or RIL, lines are usually derived from a bi-parental cross of inbred lines, so the lack of fit term can be assumed to be zero.

We note several things regarding $\boldsymbol{M}$:

1. The non-imputed values in $\boldsymbol{M}$ are generally coded as -1 corresponding to the allele inherited from parent 1, and 1 corresponding to the allele inherited from parent 2, and

2. The number of markers (r) is typically much larger than the number of lines (m)

Hence, we can write 4.7 as

$$\begin{aligned} \boldsymbol{y} &= \boldsymbol{X\tau} + \boldsymbol{Z}_g(\boldsymbol{M\alpha} + \boldsymbol{u}_e) + \boldsymbol{e}_p \\ &= \boldsymbol{X\tau} + \boldsymbol{Z}_g\boldsymbol{M\alpha} + \boldsymbol{Z}_g\boldsymbol{u}_e + \boldsymbol{e}_p \end{aligned} \tag{4.10}$$

This will be referred to as the marker (linear mixed) model. In the spirit of VanRaden (2008), we also consider the so-called line (linear mixed) model, which is given by

$$\boldsymbol{y} = \boldsymbol{X\tau} + \boldsymbol{Z}_g\boldsymbol{u}_m + \boldsymbol{Z}_g\boldsymbol{u}_e + \boldsymbol{e}_p \tag{4.11}$$

where

$$\boldsymbol{u}_m = \boldsymbol{M\alpha} \tag{4.12}$$

4.2.3 Variance models for the random genetic effects

We assume

$$\begin{bmatrix} \boldsymbol{\alpha} \\ \boldsymbol{u}_e \end{bmatrix} \sim N\left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \sigma_\alpha^2 \boldsymbol{D} & \mathbf{0} \\ \mathbf{0} & \sigma_e^2 \boldsymbol{I}_m \end{bmatrix}\right) \tag{4.13}$$

where $\boldsymbol{D} = \oplus_{i=1}^{c} \boldsymbol{D}_i$, σ_α^2 is the marker effect variance, and σ_e^2 is the residual genetic variance; c is the number of chromosomes, and $\boldsymbol{D}_i$ an $r_i \times r_i$ symmetric positive definite matrix representing the scaled variance matrix for the marker effects within chromosomes. Each chromosome has r_i markers and we note $r = \sum_{i=1}^{c} r_i$. The block diagonality form for $\boldsymbol{D}$ implicitly assumes that marker effects on different chromosomes are independent. Note that if $\boldsymbol{D} = \boldsymbol{I}_r$, that is $\boldsymbol{D}_i = \boldsymbol{I}_{r_i}$ then our model is equivalent to the marker regression model of Verbyla et al. (2012).

From 4.13, we have

$$\begin{aligned} \text{var}\left(\boldsymbol{u}_m\right) &= \sigma_\alpha^2 \boldsymbol{M}\boldsymbol{D}\boldsymbol{M}^\top = \sigma_\alpha^2 \boldsymbol{K}, \text{ and} \\ \boldsymbol{G}_g &= \text{var}\left(\boldsymbol{u}_g\right) = \sigma_\alpha^2 \boldsymbol{K} + \sigma_e^2 \boldsymbol{I}_m \end{aligned} \tag{4.14}$$

where $\boldsymbol{D}$ (and $\boldsymbol{K}$) are functions of an additional parameter ψ, which models the covariance structure of $\boldsymbol{\alpha}$. The set of genetic variance parameters are given by $\boldsymbol{\sigma}_g = (\sigma_\alpha^2, \sigma_e^2, \psi)$, and we write

$$\boldsymbol{G}_g = \boldsymbol{G}_g\left(\sigma_\alpha^2, \sigma_e^2, \psi\right)$$

to reflect the dependence of $\boldsymbol{G}_g$.

Finally, we summarise our QTL linear mixed model as

$$\boldsymbol{y} \sim N(\boldsymbol{X}\boldsymbol{\tau}, \boldsymbol{V}) \tag{4.15}$$

where

$$\begin{aligned} \boldsymbol{V} &= \boldsymbol{Z}_g \boldsymbol{G}_g \boldsymbol{Z}_g^\top + \boldsymbol{R}_p \\ &= \boldsymbol{Z}_g \boldsymbol{G}_g \boldsymbol{Z}_g^\top + \boldsymbol{Z}_p \boldsymbol{G}_p \boldsymbol{Z}_p^\top + \boldsymbol{R}_v \end{aligned} \tag{4.16}$$

where $\boldsymbol{G}_g = \boldsymbol{G}_g(\boldsymbol{\sigma}_g)$, $\boldsymbol{G}_p = \boldsymbol{G}_p(\boldsymbol{\sigma}_p)$, and $\boldsymbol{R}_v = \boldsymbol{R}_v(\boldsymbol{\sigma}_e)$. We let

$$\boldsymbol{G} = \begin{bmatrix} \boldsymbol{G}_g & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{G}_p \end{bmatrix}$$

where $\boldsymbol{\sigma}_u = \begin{bmatrix} \boldsymbol{\sigma}_g^\top & \boldsymbol{\sigma}_p^\top \end{bmatrix}^\top$ is the variance matrix for $\boldsymbol{u} = \begin{bmatrix} \boldsymbol{u}_g^\top & \boldsymbol{u}_p^\top \end{bmatrix}^\top$ with design matrix $\boldsymbol{Z} = \begin{bmatrix} \boldsymbol{Z}_g & \boldsymbol{Z}_p \end{bmatrix}$ and $\boldsymbol{R}_v(\boldsymbol{\sigma}_e)$ is the variance matrix for $\boldsymbol{e}$ dependent on the parameter vector $\boldsymbol{\sigma}_e$. These are referred to as the so-called sigma parameterization of 4.15, and the variance models in 4.15 as $\boldsymbol{G}$ structures and $\boldsymbol{R}$ structures.

4.2.4 Variance models for the marker effects

It remains to consider speciative forms for the variance matrix of $\boldsymbol{\alpha}$, where we recall

$$\begin{aligned} \text{var}\,(\boldsymbol{\alpha}) &= \sigma_\alpha^2 \boldsymbol{D} \\ &= \boldsymbol{G}_\alpha(\sigma_\alpha^2, \psi), \text{ say} \end{aligned}$$

Our development of a model for $\boldsymbol{\alpha}$ follows the aims of a QTL analysis. The primary aim of our analysis is the determination of genomic regions (or genes) that influence the expression for a quantitative trait. Implicitly, a quantitative trait is affected by many genes (or genomic regions) and our analysis how to therefore capture all regions and provide a framework for assessing the relative importance of each genomic region.

Consideration of this aim leads to the use of a geostatistical or spatial model for $\boldsymbol{\alpha}$, as the identification of genomic regions of interest is akin to the problem of spatial prediction or interpolation popularised in geo-statistics. Hence we consider the set of marker effects $\boldsymbol{\alpha}_i$ on chromosome i as being indexed by the genetic map locations taken from either a linkage map or from a physical map.

Thus, given a linkage with associated map lengths in the vector $\boldsymbol{l} = (\boldsymbol{l}_1^\top, \boldsymbol{l}_2^\top, ..., \boldsymbol{l}_c^\top)^\top$ where each vector of $\boldsymbol{l}_i$ is of r_i. Note that the first element of each $\boldsymbol{l}_i$ is arbitrarily set to 0. Given the previous partitioning of $\boldsymbol{\alpha}$ into $(\boldsymbol{\alpha}_1^\top, \dots, \boldsymbol{\alpha}_c^\top)^\top$, we write

$$\boldsymbol{\alpha}_i = \boldsymbol{\alpha}_i(\boldsymbol{l}_i)$$

for $i = 1, ..., c$, to make clear that the vector $\boldsymbol{\alpha}_i$ is indexed by the elements of $\boldsymbol{l}_i$

Our model assumes each $\boldsymbol{\alpha}_i$ is a realisation of a stationary Gaussian process with variance matrix given by $\boldsymbol{G}_{\alpha_i} = \sigma_\alpha^2 \boldsymbol{D}_i$ and let

$$\boldsymbol{G}_\alpha = \oplus_{i=1}^{c} \boldsymbol{G}_{\alpha_i}$$

Furthermore, we define the matrix $\boldsymbol{S}$ of genetic distances for linkage group i by

$$\boldsymbol{S}_i = \text{abs}\left[(\boldsymbol{l}_i \otimes \mathbf{1}_{r_i}^\top) - (\mathbf{1}_{r_i} \otimes \boldsymbol{l}_i^\top)\right] = \{S_{i;pq}\} \tag{4.17}$$

where $\mathbf{1}_{r_i}$ is a vector of length r_i comprising of 1s, and the operator abs() takes the absolute value of its matrix argument, then the elements of $\boldsymbol{D}_i$ are given by

$$\rho(S_{i;pq}, \boldsymbol{\psi}) \tag{4.18}$$

where $\rho(\cdot)$ being a correlation function with parameter vector $\boldsymbol{\psi}$, $p, q = 1, ..., r_i$, and $S_{i;pp} = 0$ and $\rho(0) = 1$.

The matrix $\boldsymbol{D}_i$ is assumed to be positive definite (and symmetric). The key characteristic and tool which describes the properties of the stochastic process $\boldsymbol{\alpha}_i(\boldsymbol{l}_i)$ is the covariance function, $\rho(\cdot)$ and its close relation, the variogram. For a stochastic process $\boldsymbol{\alpha}$ (dropping the suffix i for each of presentation), the variogram is defined to be

$$\gamma(d) = \frac{1}{2}E\left(\{\boldsymbol{\alpha}(d) - \boldsymbol{\alpha}(l-d)\}^2\right) \tag{4.19}$$

for any $d \geq 0$. Since we assume $\boldsymbol{\alpha}(\boldsymbol{l})$ is stationary, the variogram is also directly related to the (auto-)correlation function $\rho(d)$ by

$$\gamma(d) = \sigma^2_\alpha(1 - \rho(d))$$

The origins of the variogram as a descriptive test go back at least to Jowett (1952). It has been widely used in the body of methodology of geostatistics. Diggle (2002) relies heavily on the variogram for exploring covariance models in the analysis of longitudinal data. We note that there are many similarities with this application and our approach to QTL analysis, in that linkage groups can be thought of as subjects, and times can be thought of as genetic locations. The key difference however, is that genetic locations within a linkage group are directionally invariant (modulo the concept of short and long arms of the chromosome relative to the centromere).

In the following, we present a range of parametric correlation models for $\rho(d)$, which have been widely used in geostatistics, spatial statistics, and longitudinal data analysis.

Exponential:

$$\rho(d) = \exp\left(-\frac{d}{\psi}\right), \quad d \geq 0 \tag{4.20}$$

Gaussian:

$$\rho(d) = \exp\left(-\left(\frac{d}{\psi}\right)^2\right), \quad d \geq 0 \tag{4.21}$$

Spherical:

$$\rho(d) = \begin{cases} 1 - \frac{3}{2}\left(\frac{d}{\psi}\right) + \frac{1}{2}\left(\frac{d}{\psi}\right)^3, & \text{if } 0 \leq d \leq \psi \\ 0, \quad \text{if } d \geq \psi \end{cases} \tag{4.22}$$

Circular:

$$\rho(d) = \begin{cases} \frac{2}{\pi}\cos^{-1}\left(\frac{d}{\psi}\right) - \left(\frac{d}{\psi}\right)\sqrt{1 - \left(\frac{d}{\psi}\right)^2}, & \text{if } 0 \leq d \leq \psi \\ 0, \quad \text{if } d \geq \psi \end{cases} \tag{4.23}$$

Powered exponential:

$$\rho(d) = \exp\left(-\left(\frac{d}{\psi}\right)^k\right), \quad d \geq 0 \tag{4.24}$$

where k is restricted to $0 < k \leq 2$ to ensure a valid correlation function. Setting $k = 1$ or 2 gives the exponential or gaussian correlation function, respectively.

Whittle's elementary correlation:

$$\rho(d) = \frac{d}{\psi}\mathcal{K}_1\left(\frac{d}{\psi}\right), \quad d \geq 0 \tag{4.25}$$

where $\mathcal{K}_1$ is the modified Bessel function of order 1 of the third kind (Abramowitz and Stegun, 1965)

Bounded linear:

$$\rho(d) = \begin{cases} 1 - \frac{d}{\psi}, & 0 \leq d < \psi \\ 0, & d \geq \psi \end{cases} \tag{4.26}$$

Stein (1999) presents a strong case for the use of the Matérn class of correlation functions.

The isotropic Matérn correlation function is given by

$$\rho_m(d;\psi) = \left(2^{\nu-1}\Gamma(\nu)\right)^{-1}\left(\frac{d}{\psi}\right)^{\nu}\mathcal{K}_\nu\left(\frac{d}{\psi}\right) \tag{4.27}$$

where $\psi = (\phi,\nu)^\top,\ \psi > 0$, and is the so-called range parameter, $\nu > 0$ is a "smoothness" parameter, $\Gamma(\cdot)$ is the gamma function, and $\mathcal{K}_\nu(\cdot)$ is the modified Bessel function of the third kind of order ν (Abramowitz and Stegun, 1965). For a given ν (where $\nu > 0$), the range parameter, ϕ, affects the rate of decay of $\rho(\cdot)$ with increasing d. The parameter $\nu > 0$ controls the analytic smoothness of the underlying process $\boldsymbol{\alpha}(\boldsymbol{l})$.

The process $\boldsymbol{\alpha}(\boldsymbol{l})$ being $\lceil\nu\rceil - 1$ times mean-square differentiable, where $\lceil\nu\rceil$ is the smallest integer greater or equal to ν (Stein, 1999). Larger ν correspond to smoother processes.

When $\nu = h + \frac{1}{2}$ with h a non-negative integer, $\rho_m(\cdot)$ is then the product of $\exp\left(-\frac{d}{\phi}\right)$ and a polynomial of degree h in d. Thus, if $\nu = \frac{1}{2}$, then it follows

$$\rho_m(d;\phi,\frac{1}{2}) = \exp\left(-\frac{d}{\phi}\right)$$

which is the exponential correlation function. While when $\nu = 1$, then it can be shown that

$$\rho_m(d;\phi,1) = \left(-\frac{d}{\phi}\right)(K)_1\left(-\frac{d}{\phi}\right)$$

When $\nu = 1.5$, then

$$\rho_m(d;\phi,1.5) = \exp\left(-\frac{d}{\phi}\right)\left(1+\frac{d}{\phi}\right) \tag{4.28}$$

and this is the correlation function of a random (stochastic) process, $\boldsymbol{\alpha}(\boldsymbol{l})$, which is continuous and once differentiable. This form has been used by Kammann and Wand (2003). Lastly, as $\nu \to \infty$, then $\rho_m(\cdot)$ tends to the gaussian correlation function.

Thus, the Matérn correlation function offers flexibility and parsimony, and includes many

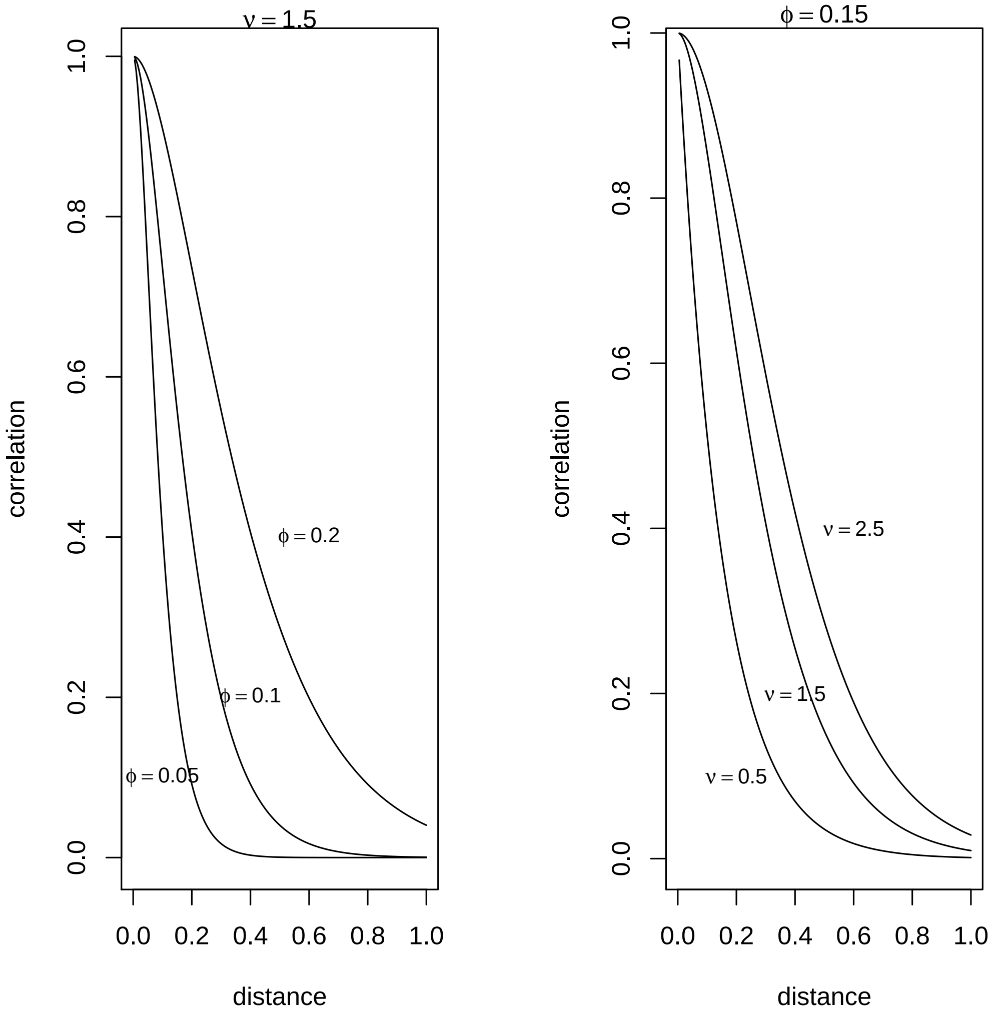


Figure 4.1: Examples of the Matérn correlation function

other correlation functions as special cases (See Figure 4.1)

4.2.5 Measurement error - The nugget effect

Our spatial QTL model for $\boldsymbol{\alpha}(\boldsymbol{l})$ does not have component which is akin to the concept on the nugget effect popularised in geostatistics, spatial statistics, and longitudinal data analysis. This idea of a nugget is to capture the error introduced by imperfect measurements. For example, in a field trial measuring soil carbon in the 0-10cm layer. Hypothetically, if one was able to remeasure at the exact same location by taking multiple soil cores, processing these in the lab, and so on, then the multiple measurements of soil carbon would be different.

Including a nugget effect into our model leads to the model for $\boldsymbol{\eta}(\boldsymbol{l})$, say

$$\boldsymbol{\eta}(\boldsymbol{l}) = \boldsymbol{\alpha}(\boldsymbol{l}) + \boldsymbol{\delta}$$

where $\boldsymbol{\delta} \sim N(\mathbf{0}, \sigma_\delta^2 \boldsymbol{I}_m)$.

5 Estimation and prediction: Exploiting the marker regression model for computational efficiency

5.1 Introduction

As mentioned in section 4, the number of lines, m, is often less than the number of markers, r. In the motivating example, $m = 142$ and $r = 1,383$, hence, we consider an alternate formulation of the two models. These are

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\tau} + \boldsymbol{Z}_g\boldsymbol{M}\boldsymbol{\alpha} + \boldsymbol{Z}_g\boldsymbol{u}_e + \boldsymbol{Z}_p\boldsymbol{u}_p + \boldsymbol{e} \tag{5.1}$$

and

$$\begin{aligned}\boldsymbol{y} &= \boldsymbol{X}\boldsymbol{\tau} + \boldsymbol{Z}_g\boldsymbol{u}_m + \boldsymbol{Z}_g\boldsymbol{u}_e + \boldsymbol{Z}_p\boldsymbol{u}_p + \boldsymbol{e} \\ &= \boldsymbol{X}\boldsymbol{\tau} + \boldsymbol{Z}\boldsymbol{u} + \boldsymbol{e} \end{aligned} \tag{5.2}$$

where $\boldsymbol{u} = (\boldsymbol{u}_m^\top, \boldsymbol{u}_e^\top, \boldsymbol{u}_p^\top)^\top$ and $\boldsymbol{Z} = \begin{bmatrix}\boldsymbol{Z}_g & \boldsymbol{Z}_{\boldsymbol{g}} & \boldsymbol{Z}_p\end{bmatrix}$. Equation 5.1 is called the marker (regression) model and equation 5.2 is the lines model.

Our approach follows those presented in Strandén and Garrick (2009). In the following, we present an approach which explots the computational efficiency of equation 5.2 for the estimation of variance parameters and develop an efficient method to predict $\boldsymbol{\alpha}$ as a post processing procedure.

5.2 Estimation and prediction

Since 5.2 is a linear mixed model, then we use Residual Maximum Likelihood (REML) (Patterson and Thompson, 1971) to estimate the variance parameters. REML estimation (i.e. maximisation of the REML log-likelihood) requires an iterative approach and we recommend the average information (AI) algorithm (Gilmour et al., 1995) as implemented in the R package `ASReml-R` (Butler et al., 2009).

Given (REML) estimates of the variance parameters, we obtain Empirical solutions for the fixed effects, $\boldsymbol{\tau}$, and Empirical Best Linear Unbiased Predictors (E-BLUP) of the random effects. From the fit of 5.2, we obtain both the E-BLUP of $\boldsymbol{u}_m$, denoted by $\tilde{\boldsymbol{u}}_m$ and the prediction error variance of $\tilde{\boldsymbol{u}}_m$, written as

$$\text{var}\left(\boldsymbol{u}_m - \tilde{\boldsymbol{u}}_m\right) = \boldsymbol{C}^{\boldsymbol{u}_m \boldsymbol{u}_m} = \text{PEV}(\tilde{\boldsymbol{u}}_m)$$

Henderson (1975) show that the PEV($\tilde{\boldsymbol{u}}_m$) is the partition of the inverse of the coefficient matrix of the MME for 5.2. These are given by

$$\boldsymbol{C}\tilde{\boldsymbol{\beta}} = \boldsymbol{W}^{\top}\boldsymbol{R}^{-1}\boldsymbol{y} \tag{5.3}$$

where $\tilde{\boldsymbol{\beta}} = \left(\hat{\boldsymbol{\tau}}^{\top}, \tilde{\boldsymbol{u}}^{\top}\right)^{\top}$, $\boldsymbol{C} = \boldsymbol{W}^{\top}\boldsymbol{R}^{-1}\boldsymbol{W} + \boldsymbol{G}^{*}$, and $\boldsymbol{G}^{*} = \begin{bmatrix} \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{G}^{-1} \end{bmatrix}$.

In the following, we illustrate how we can obtain the E-BLUP of $\boldsymbol{\alpha}$ from the E-BLUP of $\boldsymbol{u}_m$ and the PEV($\tilde{\boldsymbol{\alpha}}$) from the PEV($\tilde{\boldsymbol{u}}_m$).

5.3 Prediction of marker effects and associated prediction error variance

To simplify our proof, we rewrite the line model in the following form

$$\boldsymbol{y} = \boldsymbol{X\tau} + \boldsymbol{Z}_g\boldsymbol{u}_m + \boldsymbol{Z}_g\boldsymbol{u}_e + \boldsymbol{e}_p \tag{5.4}$$

where $\boldsymbol{e}_p = \boldsymbol{Z}_p\boldsymbol{u}_p + \boldsymbol{e}$, say. Our proof is based on the approach used by Henderson (1975) in obtaining the mixed model equations by maximisation of the so-called joint likelihood of $\boldsymbol{y}$ and $\boldsymbol{u}$ (even though it is strictly not a likelihood as $\boldsymbol{u}$ is not observed).

The joint likelihood for the line model is

$$l(\boldsymbol{y}, \boldsymbol{u}) = l(\boldsymbol{y}|\boldsymbol{u}) + l(\boldsymbol{u}) \tag{5.5}$$

but we wish to maximise equation 5.5 subject to $\boldsymbol{u}_m = \boldsymbol{M\alpha}$. Hence, we consider maximisation of the joint likelihood of $\left(\boldsymbol{y}^\top, \boldsymbol{u}^\top, \boldsymbol{\alpha}^\top\right)^\top$ subject to $\boldsymbol{u}_m = \boldsymbol{M\alpha}$.

The joint log-density of $\left(\boldsymbol{y}^\top, \boldsymbol{u}^\top, \boldsymbol{\alpha}^\top\right)^\top$ is

$$l(\boldsymbol{y}, \boldsymbol{u}, \boldsymbol{\alpha}) = l(\boldsymbol{y}|\boldsymbol{u}) + l(\boldsymbol{u}_m|\boldsymbol{\alpha}) + l(\boldsymbol{u}_e) + l(\boldsymbol{\alpha}) \tag{5.6}$$

subject to the constraint $\boldsymbol{u}_m = \boldsymbol{M\alpha}$.

Introducing Lagrangian multipliers gives

$$\begin{aligned} Q &= l(\boldsymbol{y}, \boldsymbol{u}, \boldsymbol{\alpha}) - \boldsymbol{a}^\top(\boldsymbol{u}_m - \boldsymbol{M\alpha}) \\ &= -\frac{1}{2}\left(\boldsymbol{e}_p^\top \boldsymbol{R}_p^{-1}\boldsymbol{e}_p + \boldsymbol{u}_e^\top \boldsymbol{G}_e \boldsymbol{u}_e + \boldsymbol{\alpha}^\top \boldsymbol{G}_\alpha^{-1}\boldsymbol{\alpha}\right) - \boldsymbol{a}^\top(\boldsymbol{u}_m - \boldsymbol{M\alpha}) \end{aligned} \tag{5.7}$$

say, ignoring constants, where $\boldsymbol{G}_e = \sigma_e^2\boldsymbol{I}_m$.

We maximise Q by differentiation with respect to $\boldsymbol{u}_m, \boldsymbol{\alpha}, \boldsymbol{\tau}$, and $\boldsymbol{u}_e$, and setting these quantities to zero. Thus

$$\begin{aligned}
\frac{\partial Q}{\partial \boldsymbol{u}_m} &= -\left(\boldsymbol{Z}_g^\top \boldsymbol{R}_p^{-1} \boldsymbol{Z}_g \boldsymbol{u}_m + \boldsymbol{a} + \boldsymbol{Z}_g^\top \boldsymbol{R}_p^{-1} \boldsymbol{X}\boldsymbol{\tau} - \boldsymbol{Z}_g^\top \boldsymbol{R}_p^{-1} \boldsymbol{y}\right) \\
\frac{\partial Q}{\partial \boldsymbol{a}} &= -(\boldsymbol{u}_m - \boldsymbol{M}\boldsymbol{\alpha}) \\
\frac{\partial Q}{\partial \boldsymbol{\alpha}} &= -(-\boldsymbol{M}^\top \boldsymbol{\alpha} + \boldsymbol{G}_\alpha^{-1} \boldsymbol{\alpha}) \\
\frac{\partial Q}{\partial \boldsymbol{\tau}} &= -\left(\boldsymbol{X}^\top \boldsymbol{R}_p^{-1} \boldsymbol{Z}_g \boldsymbol{u}_m + \boldsymbol{X}^\top \boldsymbol{R}_p^{-1} \boldsymbol{X}\boldsymbol{\tau} + \boldsymbol{X}^\top \boldsymbol{R}_p^{-1} \boldsymbol{Z}_g - \boldsymbol{X}^\top \boldsymbol{R}_p^{-1} \boldsymbol{y}\right) \\
\frac{\partial Q}{\partial \boldsymbol{u}_e} &= -\left(\boldsymbol{Z}_g^\top \boldsymbol{R}_p^{-1} \boldsymbol{Z}_g \boldsymbol{u}_m + (\boldsymbol{Z}_g^\top \boldsymbol{R}_p^{-1} \boldsymbol{X}\boldsymbol{\tau})\boldsymbol{u}_e + \boldsymbol{Z}_g^\top \boldsymbol{R}_p^{-1} \boldsymbol{Z}_g + \boldsymbol{G}_e^{-1} - \boldsymbol{Z}_g^\top \boldsymbol{R}_p^{-1} \boldsymbol{y}\right)
\end{aligned} \tag{5.8}$$

These differentials can be derived using the results in Appendix 7.1.3.

This, for example, we have

$$\frac{\partial Q}{\partial \boldsymbol{u}_e} = -\frac{1}{2}\left(\frac{\partial \boldsymbol{e}_p^\top \boldsymbol{R}_p^{-1} \boldsymbol{e}_p}{\partial \boldsymbol{u}_e} + \frac{\partial \boldsymbol{u}_e^\top \boldsymbol{G}_e^{-1} \boldsymbol{u}_e}{\partial \boldsymbol{u}_e}\right)$$

Now $\boldsymbol{e}_p = \boldsymbol{y} - \boldsymbol{X}\boldsymbol{\tau} - \boldsymbol{Z}_g \boldsymbol{u}_m - \boldsymbol{Z}_g \boldsymbol{u}_e$

$$\therefore \frac{\partial \boldsymbol{e}_p^\top}{\partial \boldsymbol{u}_e} = -\frac{\partial (\boldsymbol{u}_e^\top \boldsymbol{Z}_g^\top)}{\partial \boldsymbol{u}_e} = -\boldsymbol{Z}_g^\top$$

$$\begin{aligned}
\therefore -\frac{1}{2}\frac{\partial \boldsymbol{e}_p^\top \boldsymbol{R}_p^{-1} \boldsymbol{e}_p}{\partial \boldsymbol{u}_e} &= -\frac{1}{2} \times 2\boldsymbol{Z}_g^\top \boldsymbol{R}_p^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\tau} - \boldsymbol{Z}_g \boldsymbol{u}_m - \boldsymbol{Z}_g \boldsymbol{u}_e) \\
&= \boldsymbol{Z}_g^\top \boldsymbol{R}_p^{-1} \boldsymbol{y} - \boldsymbol{Z}_g^\top \boldsymbol{R}_p^{-1} \boldsymbol{X}\boldsymbol{\tau} - \boldsymbol{Z}_g^\top \boldsymbol{R}_p^{-1} \boldsymbol{Z}_g \boldsymbol{u}_m - \boldsymbol{Z}_g^\top \boldsymbol{R}_p^{-1} \boldsymbol{Z}_g \boldsymbol{u}_e
\end{aligned}$$

Also,

$$\frac{\partial \boldsymbol{u}_e^\top \boldsymbol{G}_e^{-1} \boldsymbol{u}_e}{\partial \boldsymbol{u}_e} = 2\boldsymbol{G}_e^{-1} \boldsymbol{u}_e$$

Collecting terms, we have

$$\frac{\partial Q}{\partial \boldsymbol{u}_e} = \boldsymbol{Z}_g^\top \boldsymbol{R}_p^{-1}\boldsymbol{y} - \boldsymbol{Z}_g^\top \boldsymbol{R}_p^{-1}\boldsymbol{X}\boldsymbol{\tau} - \boldsymbol{Z}_g^\top \boldsymbol{R}_p^{-1}\boldsymbol{Z}_g\boldsymbol{u}_m - \left(\boldsymbol{Z}_g^\top \boldsymbol{R}_p^{-1}\boldsymbol{Z}_g + \boldsymbol{G}_e^{-1}\right)\boldsymbol{u}_e$$

thus, proving the result.

Setting the derivatives in 5.8 to zero, we have, written as a matrix

$$\begin{bmatrix} \boldsymbol{Z}_g^\top \boldsymbol{R}_p^{-1}\boldsymbol{Z}_g & \boldsymbol{I}_m & \boldsymbol{0} & \boldsymbol{Z}_g^\top \boldsymbol{R}_p^{-1}\boldsymbol{X} & \boldsymbol{0} \\ \boldsymbol{I}_m & \boldsymbol{0} & -\boldsymbol{M} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & -\boldsymbol{M}^\top & \boldsymbol{G}_\alpha^{-1} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{X}^\top \boldsymbol{R}_p^{-1}\boldsymbol{Z}_g & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{X}^\top \boldsymbol{R}_p^{-1}\boldsymbol{X} & \boldsymbol{X}^\top \boldsymbol{R}_p^{-1}\boldsymbol{Z}_g \\ \boldsymbol{Z}_g^\top \boldsymbol{R}_p^{-1}\boldsymbol{Z}_g & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{Z}_g^\top \boldsymbol{R}_p^{-1}\boldsymbol{X} & \boldsymbol{Z}_g^\top \boldsymbol{R}_p^{-1}\boldsymbol{Z}_g + \boldsymbol{G}_e^{-1} \end{bmatrix} \begin{bmatrix} \tilde{\boldsymbol{u}}_m \\ \tilde{\boldsymbol{a}} \\ \tilde{\boldsymbol{\alpha}} \\ \hat{\boldsymbol{\tau}} \\ \tilde{\boldsymbol{u}}_e \end{bmatrix} = \begin{bmatrix} \boldsymbol{Z}_g^\top \boldsymbol{R}_p^{-1}\boldsymbol{y} \\ \boldsymbol{0} \\ \boldsymbol{0} \\ \boldsymbol{X}^\top \boldsymbol{R}_p^{-1}\boldsymbol{y} \\ \boldsymbol{Z}_g^\top \boldsymbol{R}_p^{-1}\boldsymbol{y} \end{bmatrix} \tag{5.9}$$

To simplify these equations in 5.8, we define

$$\boldsymbol{W}_0 = \begin{bmatrix} \boldsymbol{X} & \boldsymbol{Z}_g \end{bmatrix} \quad \text{and} \quad \boldsymbol{C}_0 = \boldsymbol{W}_0^\top \boldsymbol{R}_p^{-1}\boldsymbol{W}_0 + \boldsymbol{G}_0^*,$$

where

$$\boldsymbol{G}_0^* = \begin{bmatrix} \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{G}_e^{-1} \end{bmatrix}$$

Then by augmenting the coefficient matrix of 5.9 with the RHS, it can be rewritten as

$$\begin{bmatrix} \boldsymbol{y}^\top \boldsymbol{R}_p^{-1}\boldsymbol{y} & \boldsymbol{y}^\top \boldsymbol{R}_p^{-1}\boldsymbol{Z}_g & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{y}^\top \boldsymbol{R}_p^{-1}\boldsymbol{W}_0 \\ \boldsymbol{Z}_g^\top \boldsymbol{R}_p^{-1}\boldsymbol{y} & \boldsymbol{Z}_g^\top \boldsymbol{R}_p^{-1}\boldsymbol{Z}_g & \boldsymbol{I}_m & \boldsymbol{0} & \boldsymbol{Z}_g^\top \boldsymbol{R}_p^{-1}\boldsymbol{W}_0 \\ \boldsymbol{0} & \boldsymbol{I}_m & \boldsymbol{0} & -\boldsymbol{M} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & -\boldsymbol{M}^\top & \boldsymbol{G}_\alpha^{-1} & \boldsymbol{0} \\ \boldsymbol{W}_0^\top \boldsymbol{R}_p^{-1}\boldsymbol{y} & \boldsymbol{W}_0^\top \boldsymbol{R}_p^{-1}\boldsymbol{Z}_g & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{C}_0 \end{bmatrix} \tag{5.10}$$

Absorbing $(\hat{\boldsymbol{\tau}}^\top, \tilde{\boldsymbol{u}}_e^\top)^\top$ in 5.10 gives

$$\begin{aligned}
&\begin{bmatrix} \boldsymbol{y}^\top \boldsymbol{R}_p^{-1} \boldsymbol{y} & \boldsymbol{y}^\top \boldsymbol{R}_p^{-1} \boldsymbol{Z}_g & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{Z}_g^\top \boldsymbol{R}_p^{-1} \boldsymbol{y} & \boldsymbol{Z}_g^\top \boldsymbol{R}_p^{-1} \boldsymbol{Z}_g & \boldsymbol{I}_m & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I}_m & \boldsymbol{0} & -\boldsymbol{M} \\ \boldsymbol{0} & \boldsymbol{0} & -\boldsymbol{M}^\top & \boldsymbol{G}_\alpha^{-1} \end{bmatrix} \\
&- \begin{bmatrix} \boldsymbol{y}^\top \boldsymbol{R}_p^{-1} \boldsymbol{W}_0 \\ \boldsymbol{Z}_g^\top \boldsymbol{R}_p^{-1} \boldsymbol{W}_0 \\ \boldsymbol{0} \\ \boldsymbol{0} \end{bmatrix} (\boldsymbol{C}_0)^{-1} \begin{bmatrix} \boldsymbol{W}_0^\top \boldsymbol{R}_p^{-1} \boldsymbol{y} & \boldsymbol{W}_0^\top \boldsymbol{R}_p^{-1} \boldsymbol{Z}_g & \boldsymbol{0} & \boldsymbol{0} \end{bmatrix} \\
&= \begin{bmatrix} \boldsymbol{y}^\top \boldsymbol{S} \boldsymbol{y} & \boldsymbol{y}^\top \boldsymbol{S} \boldsymbol{Z}_g & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{Z}_g^\top \boldsymbol{S} \boldsymbol{y} & \boldsymbol{Z}_g^\top \boldsymbol{S} \boldsymbol{Z}_g & \boldsymbol{I}_m & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I}_m & \boldsymbol{0} & -\boldsymbol{M} \\ \boldsymbol{0} & \boldsymbol{0} & -\boldsymbol{M}^\top & \boldsymbol{G}_\alpha^{-1} \end{bmatrix}
\end{aligned} \tag{5.11}$$

which can be written as a set of reduced MMEs as

$$\begin{bmatrix} \boldsymbol{Z}_g^\top \boldsymbol{S} \boldsymbol{Z}_g & \boldsymbol{I}_m & \boldsymbol{0} \\ \boldsymbol{I}_m & \boldsymbol{0} & -\boldsymbol{M} \\ \boldsymbol{0} & -\boldsymbol{M}^\top & \boldsymbol{G}_\alpha^\top \end{bmatrix} \begin{bmatrix} \tilde{\boldsymbol{u}}_m \\ \tilde{\boldsymbol{a}} \\ \tilde{\boldsymbol{\alpha}} \end{bmatrix} = \begin{bmatrix} \boldsymbol{Z}_g^\top \boldsymbol{S} \boldsymbol{y} \\ \boldsymbol{0} \\ \boldsymbol{0} \end{bmatrix} \tag{5.12}$$

From 5.12, using the second and third rows, we have

$$\tilde{\boldsymbol{u}}_m - \boldsymbol{M} \tilde{\boldsymbol{a}} = \boldsymbol{0} \implies \tilde{\boldsymbol{u}}_m = \boldsymbol{M} \tilde{\boldsymbol{\alpha}}$$

and

$$-\boldsymbol{M}^\top \tilde{\boldsymbol{\alpha}} + \boldsymbol{G}_\alpha^{-1} \tilde{\boldsymbol{\alpha}} = \boldsymbol{0} \implies \boldsymbol{G}_\alpha^{-1} \tilde{\boldsymbol{\alpha}} = \boldsymbol{M}^\top \tilde{\boldsymbol{\alpha}} \tag{5.13}$$

Using 5.11, we now absorb $\tilde{\boldsymbol{\alpha}}$ into the equations for $\left(\tilde{\boldsymbol{u}}_m^\top, \tilde{\boldsymbol{\alpha}}^\top\right)^\top$ as follows

$$\begin{aligned}
&\begin{bmatrix} \boldsymbol{y}^\top\boldsymbol{S}\boldsymbol{y} & \boldsymbol{y}^\top\boldsymbol{S}\boldsymbol{Z}_g & \mathbf{0} \\ \boldsymbol{Z}_g^\top\boldsymbol{S}\boldsymbol{y} & \boldsymbol{Z}_g^\top\boldsymbol{S}\boldsymbol{Z}_g & \boldsymbol{I}_m \\ \mathbf{0} & \boldsymbol{I}_m & \mathbf{0} \end{bmatrix} - \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \\ -\boldsymbol{M} \end{bmatrix} (\boldsymbol{G}_\alpha) \begin{bmatrix} \mathbf{0} & \mathbf{0} & -\boldsymbol{M}^\top \end{bmatrix} \\
&\qquad = \begin{bmatrix} \boldsymbol{y}^\top\boldsymbol{S}\boldsymbol{y} & \boldsymbol{y}^\top\boldsymbol{S}\boldsymbol{Z}_g & \mathbf{0} \\ \boldsymbol{Z}_g^\top\boldsymbol{S}\boldsymbol{y} & \boldsymbol{Z}_g^\top\boldsymbol{S}\boldsymbol{Z}_g & \boldsymbol{I}_m \\ \mathbf{0} & \boldsymbol{I}_m & -\boldsymbol{G}_m \end{bmatrix}
\end{aligned} \tag{5.14}$$

where $\boldsymbol{G}_m = \boldsymbol{M}\boldsymbol{G}_\alpha\boldsymbol{M}^\top = \sigma_\alpha^2\boldsymbol{M}\boldsymbol{D}\boldsymbol{M}^\top = \sigma_\alpha^2\boldsymbol{K}$.

We can write 5.14 as a set of reduced MMEs for $(\tilde{\boldsymbol{u}}_m^\top, \tilde{\boldsymbol{a}}^\top)^\top$ as

$$\begin{bmatrix} \boldsymbol{Z}_g^\top\boldsymbol{S}\boldsymbol{Z}_g & \boldsymbol{I}_m \\ \boldsymbol{I}_m & -\boldsymbol{G}_m \end{bmatrix} \begin{bmatrix} \tilde{\boldsymbol{u}}_m \\ \tilde{\boldsymbol{a}} \end{bmatrix} = \begin{bmatrix} \boldsymbol{Z}_g^\top\boldsymbol{S}\boldsymbol{y} \\ \mathbf{0} \end{bmatrix} \tag{5.15}$$

From the second equation in 5.15, we have

$$\tilde{\boldsymbol{u}}_m - \boldsymbol{G}_m\tilde{\boldsymbol{a}} = 0 \implies \tilde{\boldsymbol{u}}_m = \boldsymbol{G}_m^{-1}\tilde{\boldsymbol{a}} \tag{5.16}$$

Finally, absorbing the equation for $\tilde{\boldsymbol{a}}$ into the equation for $\tilde{\boldsymbol{u}}_m$, we have

$$\begin{aligned}
&\begin{bmatrix} \boldsymbol{y}^\top\boldsymbol{S}\boldsymbol{y} & \boldsymbol{y}^\top\boldsymbol{S}\boldsymbol{Z}_g \\ \boldsymbol{Z}_g^\top\boldsymbol{S}\boldsymbol{y} & \boldsymbol{Z}_g^\top\boldsymbol{S}\boldsymbol{Z}_g \end{bmatrix} - \begin{bmatrix} \mathbf{0} \\ \boldsymbol{I}_m \end{bmatrix} (-\boldsymbol{G}_m)^{-1} \begin{bmatrix} \mathbf{0} & \boldsymbol{I}_m \end{bmatrix} \\
&\qquad = \begin{bmatrix} \boldsymbol{y}^\top\boldsymbol{S}\boldsymbol{y} & \boldsymbol{y}^\top\boldsymbol{S}\boldsymbol{Z}_g \\ \boldsymbol{Z}_g^\top\boldsymbol{S}\boldsymbol{y} & \boldsymbol{Z}_g^\top\boldsymbol{S}\boldsymbol{Z}_g + \boldsymbol{G}_m^{-1} \end{bmatrix}
\end{aligned}$$

and hence

$$(\boldsymbol{Z}_g^\top \boldsymbol{S}\boldsymbol{Z}_g + \boldsymbol{G}_m^{-1})\tilde{\boldsymbol{u}}_m = \boldsymbol{Z}_g^\top \boldsymbol{S}\boldsymbol{y} \tag{5.17}$$

Hence, it follows that

$$\text{PEV}(\tilde{\boldsymbol{u}}_m) = (\boldsymbol{Z}_g^\top \boldsymbol{S}\boldsymbol{Z}_g + \boldsymbol{G}_m^{-1})^{-1} \tag{5.18}$$

Using 5.16 and 5.13, we have

$$\boldsymbol{G}_\alpha^{-1}\tilde{\boldsymbol{\alpha}} = \boldsymbol{M}^\top \tilde{\boldsymbol{a}} \quad \text{and} \quad \tilde{\boldsymbol{a}} = \boldsymbol{G}_m^{-1}\tilde{\boldsymbol{u}}_m$$

so

$$\begin{aligned}\tilde{\boldsymbol{\alpha}} &= \boldsymbol{G}_\alpha \boldsymbol{M}^\top \boldsymbol{G}_m^{-1}\tilde{\boldsymbol{u}_m} \\ &= \boldsymbol{D}\boldsymbol{M}^\top \boldsymbol{K}^{-1}\tilde{\boldsymbol{u}}_m \end{aligned} \tag{5.19}$$

Recapping, we have the following key results

$$\begin{aligned}\boldsymbol{C}_1\tilde{\boldsymbol{\beta}}_1 &= \boldsymbol{W}_1^\top \boldsymbol{S}\boldsymbol{y} \\ \boldsymbol{C}_2\tilde{\boldsymbol{\beta}}_2 &= \boldsymbol{W}_2^\top \boldsymbol{S}\boldsymbol{y} \\ \boldsymbol{C}_3\tilde{\boldsymbol{\beta}}_3 &= \boldsymbol{W}_3^\top \boldsymbol{S}\boldsymbol{y}\end{aligned} \tag{5.20}$$

where

$$
\begin{aligned}
\boldsymbol{C}_1 &= \begin{bmatrix} \boldsymbol{Z}_g^\top \boldsymbol{S} \boldsymbol{Z}_g & \boldsymbol{I}_m & \mathbf{0} \\ \boldsymbol{I}_m & \mathbf{0} & -\boldsymbol{M} \\ \mathbf{0} & -\boldsymbol{M}^\top & \boldsymbol{G}_\alpha^\top \end{bmatrix} \\
\tilde{\boldsymbol{\beta}}_1 &= \begin{bmatrix} \tilde{\boldsymbol{u}}_m \\ \tilde{\boldsymbol{a}} \\ \tilde{\boldsymbol{\alpha}} \end{bmatrix} \\
\boldsymbol{W}_1 &= \begin{bmatrix} \boldsymbol{Z}_g & \mathbf{0} & \mathbf{0} \end{bmatrix} \\
\boldsymbol{C}_2 &= \begin{bmatrix} \boldsymbol{Z}_g^\top \boldsymbol{S} \boldsymbol{Z}_g & \boldsymbol{I}_m \\ \boldsymbol{I}_m & -\boldsymbol{G}_m \end{bmatrix} \\
\tilde{\boldsymbol{\beta}}_2 &= \begin{bmatrix} \tilde{\boldsymbol{u}}_m \\ \tilde{\boldsymbol{a}} \end{bmatrix} \\
\boldsymbol{W}_2 &= \begin{bmatrix} \boldsymbol{Z}_g & \mathbf{0} \end{bmatrix} \\
\boldsymbol{C}_3 &= (\boldsymbol{Z}_g^\top \boldsymbol{S} \boldsymbol{Z}_g + \boldsymbol{G}_m^{-1}) \\
\tilde{\boldsymbol{\beta}}_3 &= \tilde{\boldsymbol{u}}_m \\
\boldsymbol{W}_3 &= \boldsymbol{Z}_g
\end{aligned}
$$

Now it follows that the PEV of any $\tilde{\boldsymbol{\beta}}_i$ is given by $\boldsymbol{C}_i^{-1},\ i = 1, 2, 3$ (Henderson, 1975).

Thus, we have, for example $i = 3$

$$
\begin{aligned}
\mathrm{PEV}(\tilde{\boldsymbol{u}}_m) = \mathrm{PEV}(\tilde{\boldsymbol{\beta}}_3) &= \boldsymbol{C}_3^{-1} \\
&= (\boldsymbol{Z}_g^\top \boldsymbol{S} \boldsymbol{Z}_g + \boldsymbol{G}_m^{-1})^{-1} \quad \text{(see 5.17)}
\end{aligned}
$$

Furthermore, $i = 1$, then

$$\begin{aligned}\mathrm{PEV}(\tilde{\boldsymbol{\beta}}_1) &= \boldsymbol{C}_1^{-1} \\ &= \begin{bmatrix} \boldsymbol{C}_1^{\boldsymbol{u}_m\boldsymbol{u}_m} & \boldsymbol{C}_1^{\boldsymbol{u}_m\boldsymbol{a}} & \boldsymbol{C}_1^{\boldsymbol{u}_m\boldsymbol{\alpha}} \\ \boldsymbol{C}_1^{\boldsymbol{a}\boldsymbol{u}_m} & \boldsymbol{C}_1^{\boldsymbol{a}\boldsymbol{a}} & \boldsymbol{C}_1^{\boldsymbol{a}\boldsymbol{\alpha}} \\ \boldsymbol{C}_1^{\boldsymbol{\alpha}\boldsymbol{u}_m} & \boldsymbol{C}_1^{\boldsymbol{\alpha}\boldsymbol{a}} & \boldsymbol{C}_1^{\boldsymbol{\alpha}\boldsymbol{\alpha}} \end{bmatrix}\end{aligned} \tag{5.21}$$

Before we proceed, we present a result on the inverse of a partitioned matrix

Result 6.3.1 (Searle, 1928)

Let $\boldsymbol{C} = \begin{bmatrix} \boldsymbol{C}_{11} & \boldsymbol{C}_{12} \\ \boldsymbol{C}_{21} & \boldsymbol{C}_{22} \end{bmatrix}$ be a positive definite symmetric matrix, then

$$\boldsymbol{C} = \begin{bmatrix} \boldsymbol{C}_{11} & \boldsymbol{C}_{12} \\ \boldsymbol{C}_{21} & \boldsymbol{C}_{22} \end{bmatrix} = \begin{bmatrix} \boldsymbol{T} & -\boldsymbol{T}\boldsymbol{U}^\top \\ -\boldsymbol{U}\boldsymbol{T} & \boldsymbol{C}_{22}^{-1} + \boldsymbol{U}\boldsymbol{T}\boldsymbol{U}^\top \end{bmatrix}$$

where $\boldsymbol{T} = (\boldsymbol{C}_{11} - \boldsymbol{C}_{12}\boldsymbol{C}_{22}^{-1}\boldsymbol{C}_{21})^{-1}$, and $\boldsymbol{U} = \boldsymbol{C}_{22}^{-1}\boldsymbol{C}_{21}$.

Now, we use Result 6.3.1 several times in the following. Firstly, let $\boldsymbol{C} = \boldsymbol{C}_1$, and consider

$$\boldsymbol{C} = \begin{bmatrix} \boldsymbol{Z}_g^\top \boldsymbol{S}\boldsymbol{Z}_g & \boldsymbol{I}_m & \boldsymbol{0} \\ \boldsymbol{I}_m & \boldsymbol{0} & -\boldsymbol{M} \\ \boldsymbol{0} & -\boldsymbol{M}^\top & \boldsymbol{G}_\alpha^{-1} \end{bmatrix} = \begin{bmatrix} \boldsymbol{C}_{11} & \boldsymbol{C}_{12} \\ \boldsymbol{C}_{21} & \boldsymbol{C}_{22} \end{bmatrix}$$

where $\boldsymbol{C}_{11} = \begin{bmatrix} \boldsymbol{Z}_g^\top \boldsymbol{S}\boldsymbol{Z}_g & \boldsymbol{I}_m \\ \boldsymbol{I}_m & \boldsymbol{0} \end{bmatrix}$, $\boldsymbol{C}_{12} = \begin{bmatrix} \boldsymbol{0} \\ -\boldsymbol{M} \end{bmatrix}$, $\boldsymbol{C}_{21} = (\boldsymbol{C}_{12})^\top$ and $\boldsymbol{C}_{22} = \boldsymbol{G}_\alpha^{-1}$.

Hence, the PEV($\tilde{\boldsymbol{\alpha}}$) is given by

$$\begin{aligned} \text{PEV}(\tilde{\boldsymbol{\alpha}}) &= \boldsymbol{C}^{22} \\ &= \boldsymbol{C}_{22}^{-1} + \boldsymbol{U}\boldsymbol{T}\boldsymbol{U}^{\top} \qquad (5.22) \\ & \qquad (5.23) \end{aligned}$$

where $\boldsymbol{U} = \boldsymbol{C}_{22}^{-1}\boldsymbol{C}_{21}$, and

$$\begin{aligned} \boldsymbol{T} &= (\boldsymbol{C}_{11} - \boldsymbol{C}_{12}\boldsymbol{C}_{22}^{-1}\boldsymbol{C}_{21})^{-1} \\ &= \begin{bmatrix} \boldsymbol{T}^{11} & \boldsymbol{T}^{12} \\ \boldsymbol{T}^{21} & \boldsymbol{T}^{22} \end{bmatrix} \end{aligned}$$

Thus, from 5.23, we have

$$\begin{aligned} \text{PEV}(\tilde{\boldsymbol{\alpha}}) &= \boldsymbol{C}_{22}^{-1} + \boldsymbol{C}_{22}^{-1} \begin{bmatrix} \mathbf{0} & -\boldsymbol{M}^{\top} \end{bmatrix} \begin{bmatrix} \boldsymbol{T}^{11} & \boldsymbol{T}^{12} \\ \boldsymbol{T}^{21} & \boldsymbol{T}^{22} \end{bmatrix} \begin{bmatrix} \mathbf{0} \\ -\boldsymbol{M} \end{bmatrix} \boldsymbol{C}_{22}^{-1} \\ &= \boldsymbol{C}_{22}^{-1} + \boldsymbol{C}_{22}^{-1}\boldsymbol{M}^{\top}\boldsymbol{T}^{22}\boldsymbol{M}\boldsymbol{C}_{22}^{-1} \end{aligned}$$

Since $\boldsymbol{C}_{22}^{-1} = \boldsymbol{G}_{\alpha}$, then

$$\implies \text{PEV}(\tilde{\boldsymbol{\alpha}}) = \boldsymbol{G}_{\alpha} + \boldsymbol{G}_{\alpha}\boldsymbol{M}^{\top}\boldsymbol{T}^{22}\boldsymbol{M}\boldsymbol{G}_{\alpha} \qquad (5.24)$$

To obtain $\boldsymbol{T}^{22}$ in 5.24, we set $\boldsymbol{C} = \boldsymbol{C}_2$, where $\boldsymbol{C}_2 = \begin{bmatrix} \boldsymbol{Z}_g^{\top}\boldsymbol{S}\boldsymbol{Z}_g & \boldsymbol{I}_m \\ \boldsymbol{I}_m & -\boldsymbol{G}_m \end{bmatrix} = \boldsymbol{T}^{-1}$

$$\begin{aligned} \implies \boldsymbol{T}^{22} &= -\boldsymbol{G}_m^{-1} + \boldsymbol{G}_m^{-1}\boldsymbol{C}^{\boldsymbol{u}_m\boldsymbol{u}_m}\boldsymbol{G}_m^{-1} \\ &= \boldsymbol{G}_m^{-1}\boldsymbol{C}^{\boldsymbol{u}_m\boldsymbol{u}_m}\boldsymbol{G}_m^{-1} - \boldsymbol{G}_m^{-1} \qquad (5.25) \end{aligned}$$

as $\boldsymbol{C}^{\boldsymbol{u}_m\boldsymbol{u}_m} = \text{PEV}(\tilde{\boldsymbol{u}}_m) = (\boldsymbol{Z}_g^\top \boldsymbol{S}\boldsymbol{Z}_g + \boldsymbol{G}_m^{-1})^{-1}$

Thus, using 5.25 and 5.24, we have

$$\begin{aligned}
\text{PEV}(\tilde{\boldsymbol{\alpha}}) &= \boldsymbol{G}_\alpha + \boldsymbol{G}_\alpha \boldsymbol{M}^\top \left(\boldsymbol{G}_m^{-1}\boldsymbol{C}^{\boldsymbol{u}_m\boldsymbol{u}_m}\boldsymbol{G}_m^{-1} - \boldsymbol{G}_m^{-1}\right)\boldsymbol{M}\boldsymbol{G}_\alpha \\
&= \boldsymbol{G}_\alpha - \boldsymbol{G}_\alpha \boldsymbol{M}^\top \boldsymbol{G}_m^{-1}\boldsymbol{M}\boldsymbol{G}_\alpha + \boldsymbol{G}_\alpha \boldsymbol{M}^\top \boldsymbol{G}_m^{-1}\boldsymbol{C}^{\boldsymbol{u}_m\boldsymbol{u}_m}\boldsymbol{G}_m^{-1}\boldsymbol{M}\boldsymbol{G}_\alpha
\end{aligned} \tag{5.26}$$

Noting that $\boldsymbol{G}_\alpha = \sigma_\alpha^2\boldsymbol{D}$ and $\boldsymbol{G}_m = \sigma_\alpha^2\boldsymbol{K}$, for $\boldsymbol{K} = \boldsymbol{M}\boldsymbol{D}\boldsymbol{M}^\top$, then

$$\begin{aligned}
\text{PEV}(\tilde{\boldsymbol{\alpha}}) &= \sigma_\alpha^2\boldsymbol{D} - \sigma_\alpha^2\boldsymbol{D}\boldsymbol{M}^\top(\sigma_\alpha^2)^{-1}(\boldsymbol{K})^{-1}\boldsymbol{M}\sigma_\alpha^2\boldsymbol{D} \\
&\qquad + \sigma_\alpha^2\boldsymbol{D}\boldsymbol{M}^\top(\sigma_\alpha^2)^{-1}(\boldsymbol{K})^{-1}\boldsymbol{C}^{\boldsymbol{u}_m\boldsymbol{u}_m}(\sigma_\alpha^2)^{-1}(\boldsymbol{K})^{-1}\boldsymbol{M}\sigma_\alpha^2\boldsymbol{D} \\
&= \sigma_\alpha^2(\boldsymbol{D} - \boldsymbol{D}\boldsymbol{M}^\top\boldsymbol{K}^{-1}\boldsymbol{M}\boldsymbol{D}) + \boldsymbol{D}\boldsymbol{M}^\top\boldsymbol{K}^{-1}\boldsymbol{C}^{\boldsymbol{u}_m\boldsymbol{u}_m}\boldsymbol{K}^{-1}\boldsymbol{M}\boldsymbol{D} \\
&= \sigma_\alpha^2\boldsymbol{D.} + \boldsymbol{D}\boldsymbol{M}^\top\boldsymbol{K}^{-1}\text{PEV}(\tilde{\boldsymbol{u}}_m)\boldsymbol{K}^{-1}\boldsymbol{M}\boldsymbol{D}
\end{aligned} \tag{5.27}$$

where $\boldsymbol{D.} = \boldsymbol{D} - \boldsymbol{D}\boldsymbol{M}^\top\boldsymbol{K}^{-1}\boldsymbol{M}\boldsymbol{D}$ □

6 Conclusion

The need for a consolidated framework for the determination of QTLs in plants and animals which is aligned with many of the current approaches used in genomic selection is a key issue. In this project we have developed the framework and an efficient approach to estimation and prediction even when $m \ll r$. An increased efficiency of the approach will hopefully not only make the analysis faster but also encourage the adoption of a fully efficient one-stage approach to QTL analysis to replace the more popular two-stage approaches. Since publication, Verbyla et al. (2007) has received only 57 citations (Google Scholar), whilst WGAIM (Taylor and Verbyla, 2011) and RWGAIM (Verbyla et al., 2012) have only been cited 15 and 16 times, respectively. Possible reasons for the lack of adoption of these techniques could be the need for a recursive approach for QTL detection where often large and complex linear mixed models have to be fitted many times. Furthermore, the use of the alternate outlier model cannot be implemented as a score test for non-gaussian data. For example, Boden et al. (2015) were forced to resort to the use of Wald tests using a marker genome scan. This approach was computationally intensive and piecemeal.

Our new approach requires only one linear mixed model to be fitted, leaving the identification of the location of QTLs as an efficient post-processing method. There remains much work to be done, however, before it can be recommended as an alternative to the other so-called whole-genome approaches. For example, a range of model diagnostics need to be developed and tested. The use of the sample variogram is an obvious candidate.

Secondly, an approach to determine coverage intervals for putative QTLs needs to be developed, and it is likely that exact methods would be suitable for this.

Thirdly, it would be useful to undertake a simulation study comparing the performance of the current method with other whole genome approaches. Some preliminary empirical results suggest that our innovative approach produces comparable results to WGAIM traits with a simple underlying genetic model, though substantial discrepancies occur for the analysis of more complex traits.

7 Appendix

7.1 Matrix results

7.1.1 Matrix absorption

Let $\boldsymbol{cb} = \boldsymbol{r}$ be an equation for $\boldsymbol{b} = \begin{bmatrix} \boldsymbol{b}_1 \\ \boldsymbol{b}_2 \end{bmatrix}$, that is

$$\begin{bmatrix} \boldsymbol{C}_{11} & \boldsymbol{C}_{12} \\ \boldsymbol{C}_{21} & \boldsymbol{C}_{22} \end{bmatrix} \begin{bmatrix} \boldsymbol{b}_1 \\ \boldsymbol{b}_2 \end{bmatrix} = \begin{bmatrix} \boldsymbol{r}_1 \\ \boldsymbol{r}_2 \end{bmatrix}$$

$$\Longrightarrow \boldsymbol{C}_{11}\boldsymbol{b}_1 + \boldsymbol{C}_{12}\boldsymbol{b}_2 = \boldsymbol{r}_1 \tag{7.1}$$

$$\boldsymbol{C}_{21}\boldsymbol{b}_1 + \boldsymbol{C}_{22}\boldsymbol{b}_2 = \boldsymbol{r}_2 \tag{7.2}$$

From equation 7.2, we have

$$\boldsymbol{C}_{22}\boldsymbol{b}_2 = \boldsymbol{r}_2 - \boldsymbol{C}_{21}\boldsymbol{b}_1$$

$$\Longrightarrow \boldsymbol{b}_2 = \boldsymbol{C}_{22}^{-1}(\boldsymbol{r}_2 - \boldsymbol{C}_{21}\boldsymbol{b}_1) \tag{7.3}$$

substituting 7.3 into 7.1, we get

$$\begin{aligned}
&\boldsymbol{C}_{11}\boldsymbol{b}_1 + \boldsymbol{C}_{12}\boldsymbol{C}_{22}^{-1}(\boldsymbol{r}_2 - \boldsymbol{C}_{21}\boldsymbol{b}_1) = \boldsymbol{r}_1 \\
\Longrightarrow &(\boldsymbol{C}_{11} - \boldsymbol{C}_{12}\boldsymbol{C}_{22}^{-1}\boldsymbol{C}_{21})\boldsymbol{b}_1 = \boldsymbol{r}_1 - \boldsymbol{C}_{12}\boldsymbol{C}_{22}^{-1}\boldsymbol{r}_2 \qquad (7.4)
\end{aligned}$$

that is if $\boldsymbol{C}_{11} = \boldsymbol{X}_1^\top \boldsymbol{X}_1$, $\boldsymbol{C}_{12} = \boldsymbol{X}_1^\top \boldsymbol{X}_2$, $\boldsymbol{C}_{21} = \boldsymbol{X}_2^\top \boldsymbol{X}_1$, $\boldsymbol{C}_{22} = \boldsymbol{X}_2^\top \boldsymbol{X}_2$, and $\boldsymbol{r}_1 = \boldsymbol{X}_1^\top \boldsymbol{y}$, $\boldsymbol{r}_2 = \boldsymbol{X}_2^\top \boldsymbol{y}$, then we write

$$\begin{bmatrix} \boldsymbol{y}^\top\boldsymbol{y} & \boldsymbol{y}^\top\boldsymbol{X}_1 & \boldsymbol{y}^\top\boldsymbol{X}_2 \\ \boldsymbol{X}_1^\top\boldsymbol{y} & \boldsymbol{X}_1^\top\boldsymbol{X}_1 & \boldsymbol{X}_1^\top\boldsymbol{X}_2 \\ \boldsymbol{X}_2^\top\boldsymbol{y} & \boldsymbol{X}_2^\top\boldsymbol{X}_1 & \boldsymbol{X}_2^\top\boldsymbol{X}_2 \end{bmatrix} = \begin{bmatrix} \boldsymbol{y}^\top\boldsymbol{y} & \boldsymbol{r}_1^\top & \boldsymbol{r}_2^\top \\ \boldsymbol{r}_1 & \boldsymbol{C}_{11} & \boldsymbol{C}_{12} \\ \boldsymbol{r}_2 & \boldsymbol{C}_{21} & \boldsymbol{C}_{22} \end{bmatrix} \qquad (7.5)$$

then the absorption is the matrix operation which reduces 7.5 to

$$\begin{aligned}
&\begin{bmatrix} \boldsymbol{y}^\top\boldsymbol{y} & \boldsymbol{r}_1^\top \\ \boldsymbol{r}_1 & \boldsymbol{C}_{11} \end{bmatrix} - \begin{bmatrix} \boldsymbol{r}_2^\top \\ \boldsymbol{C}_{12} \end{bmatrix} \boldsymbol{C}_{22}^{-1} \begin{bmatrix} \boldsymbol{r}_2 & \boldsymbol{C}_{21} \end{bmatrix} \\
= &\begin{bmatrix} \boldsymbol{y}^\top\boldsymbol{y} & \boldsymbol{r}_1^\top \\ \boldsymbol{r}_1 & \boldsymbol{C}_{11} \end{bmatrix} - \begin{bmatrix} \boldsymbol{r}_2^\top\boldsymbol{C}_{22}^{-1}\boldsymbol{r}_2 & \boldsymbol{r}_2^\top\boldsymbol{C}_{22}^{-1}\boldsymbol{C}_{21} \\ \boldsymbol{C}_{12}\boldsymbol{C}_{22}^{-1}\boldsymbol{r}_2 & \boldsymbol{C}_{12}\boldsymbol{C}_{22}^{-1}\boldsymbol{C}_{21} \end{bmatrix} \\
= &\begin{bmatrix} \boldsymbol{y}^\top\boldsymbol{y} - \boldsymbol{r}_2^\top\boldsymbol{C}_{22}^{-1}\boldsymbol{r}_2 & \boldsymbol{r}_1^\top - \boldsymbol{r}_2^\top\boldsymbol{C}_{22}^{-1}\boldsymbol{C}_{21} \\ \boldsymbol{r}_1 - \boldsymbol{C}_{12}\boldsymbol{C}_{22}^{-1}\boldsymbol{r}_2 & \boldsymbol{C}_{11} - \boldsymbol{C}_{12}\boldsymbol{C}_{22}^{-1}\boldsymbol{C}_{21} \end{bmatrix} \qquad (7.6)
\end{aligned}$$

The second row of equation 7.6 can be considered as the reduced equation for $\boldsymbol{b}_1$, that is,

$$(\boldsymbol{C}_{11} - \boldsymbol{C}_{12}\boldsymbol{C}_{22}^{-1}\boldsymbol{C}_{21})\boldsymbol{b}_1 = \boldsymbol{r}_1 - \boldsymbol{C}_{12}\boldsymbol{C}_{22}^{-1}\boldsymbol{r}_2$$

which is equivalent to equation 7.4. The element in row 1 of column 1 is the total SS (of

$\boldsymbol{y}$) adjusted for $\boldsymbol{X}_2$.

7.1.2 Inverses of partitioned matrices

Using the example by Searle (1928), Observe that for non-singular matrices $\boldsymbol{R}$ and $\boldsymbol{S}$,

$$\begin{bmatrix} \boldsymbol{R} & 0 \\ \boldsymbol{X} & \boldsymbol{S} \end{bmatrix}^{-1} = \begin{bmatrix} \boldsymbol{R}^{-1} & 0 \\ \boldsymbol{S}^{-1}\boldsymbol{X}\boldsymbol{R}^{-1} & \boldsymbol{S}^{-1} \end{bmatrix} \tag{7.7}$$

and also consider the partitioned matrix,

$$\begin{bmatrix} \boldsymbol{A} & \boldsymbol{B} \\ \boldsymbol{C} & \boldsymbol{D} \end{bmatrix} = \begin{bmatrix} \boldsymbol{A} & 0 \\ \boldsymbol{C} & \boldsymbol{D}-\boldsymbol{C}\boldsymbol{A}^{-1}\boldsymbol{B} \end{bmatrix} \begin{bmatrix} \boldsymbol{I} & \boldsymbol{A}^{-1}\boldsymbol{B} \\ 0 & \boldsymbol{I} \end{bmatrix} \tag{7.8}$$

Using the result from 7.7 and its transpose on 7.8, we get

$$\begin{bmatrix} \boldsymbol{A} & \boldsymbol{B} \\ \boldsymbol{C} & \boldsymbol{D} \end{bmatrix}^{-1} = \begin{bmatrix} \boldsymbol{I} & -\boldsymbol{A}^{-1}\boldsymbol{B} \\ 0 & \boldsymbol{I} \end{bmatrix} \begin{bmatrix} \boldsymbol{A}^{-1} & 0 \\ -(\boldsymbol{D}-\boldsymbol{C}\boldsymbol{A}^{-1}\boldsymbol{B})^{-1}\boldsymbol{C}\boldsymbol{A}^{-1} & (\boldsymbol{D}-\boldsymbol{C}\boldsymbol{A}^{-1}\boldsymbol{B})^{-1} \end{bmatrix} \tag{7.9}$$

which can be rewritten as

$$\begin{bmatrix} \boldsymbol{A} & \boldsymbol{B} \\ \boldsymbol{C} & \boldsymbol{D} \end{bmatrix}^{-1} = \begin{bmatrix} \boldsymbol{A}^{-1} & 0 \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} -\boldsymbol{A}^{-1}\boldsymbol{B} \\ \boldsymbol{I} \end{bmatrix} (\boldsymbol{D}-\boldsymbol{C}\boldsymbol{A}^{-1}\boldsymbol{B})^{-1} \begin{bmatrix} -\boldsymbol{C}\boldsymbol{A}^{-1} & \boldsymbol{I} \end{bmatrix} \tag{7.10}$$

7.1.3 Results of Matrix Differentiation

Some general results regarding derivatives. Let $\boldsymbol{x}$ be an $n \times 1$ vector, $\boldsymbol{A}$ be an $n \times p$ matrix, and $\boldsymbol{G}$ be an $n \times n$ symmetric matrix. Then

1. $\frac{\partial}{\partial \boldsymbol{x}}\boldsymbol{x}^\top \boldsymbol{A} = \boldsymbol{A}$

where, for example, if λ is any scalar, then the differential with respect to a scalar is defined to be

$$\frac{\partial \lambda}{\partial \boldsymbol{x}} = \begin{bmatrix} \frac{\partial \lambda}{\partial x_1} \\ \frac{\partial \lambda}{\partial x_2} \\ \vdots \\ \frac{\partial \lambda}{\partial x_n} \end{bmatrix}$$

Further, if $\boldsymbol{y}^\top = \begin{bmatrix} y_1 & \dots & y_n \end{bmatrix} = \begin{bmatrix} \boldsymbol{x}^\top \boldsymbol{a}_1 & \dots & \boldsymbol{x}^\top \boldsymbol{a}_n \end{bmatrix}$, then

$$\begin{aligned} \frac{\partial \boldsymbol{y}^\top}{\partial \boldsymbol{x}} &= \begin{bmatrix} \frac{\partial y_1}{\partial \boldsymbol{x}} & \frac{\partial y_2}{\partial \boldsymbol{x}} & \frac{\partial y_3}{\partial \boldsymbol{x}} & \dots & \frac{\partial y_n}{\partial \boldsymbol{x}} \end{bmatrix} \\ &= \begin{bmatrix} \frac{\partial \boldsymbol{x}^\top \boldsymbol{a}_1}{\partial \boldsymbol{x}} & \frac{\partial \boldsymbol{x}^\top \boldsymbol{a}_2}{\partial \boldsymbol{x}} & \frac{\partial \boldsymbol{x}^\top \boldsymbol{a}_3}{\partial \boldsymbol{x}} & \dots & \frac{\partial \boldsymbol{x}^\top \boldsymbol{a}_n}{\partial \boldsymbol{x}} \end{bmatrix} \\ &= \begin{bmatrix} \boldsymbol{a}_1 & \boldsymbol{a}_2 & \boldsymbol{a}_3 & \dots & \boldsymbol{a}_n \end{bmatrix} = \boldsymbol{A} \end{aligned}$$

Thus $\frac{\partial \boldsymbol{x}^\top \boldsymbol{A}}{\partial \boldsymbol{x}} = \boldsymbol{A}$

2. $\frac{\partial \boldsymbol{A}\boldsymbol{x}}{\partial \boldsymbol{x}} = \boldsymbol{A}^\top$

3. $\frac{\partial \boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x}}{\partial \boldsymbol{x}} = 2\boldsymbol{A}\boldsymbol{x}$

4. If $\boldsymbol{\alpha}$ is a $p \times 1$ vector, then $\frac{\partial \boldsymbol{x}^\top \boldsymbol{G}\boldsymbol{x}}{\partial \boldsymbol{\alpha}} = 2\frac{\partial \boldsymbol{x}^\top}{\partial \boldsymbol{\alpha}}\boldsymbol{G}\boldsymbol{x}$

7.2 Kosambi Distance Function

The Kosambi (Kosambi, 1944) map function that describes the genetic distance d is also related to the recombination fraction θ can be shown as

$$\begin{aligned}
d &= \frac{1}{4} ln \left(\frac{1+2\theta}{1-2\theta} \right) \\
e^{4d} &= \frac{1+2\theta}{1-2\theta} \\
e^{-4d} &= \frac{1-2\theta}{1+2\theta} \\
1-2\theta &= (1+2\theta)e^{-4d} \\
1-2\theta &= e^{-4d} + 2\theta e^{-4d} \\
1-e^{-4d} &= 2\theta + 2\theta e^{-4d} \\
\theta &= \frac{1}{2} \left(\frac{1-e^{-4d}}{1+e^{-4d}} \right) \\
M_K(d) &= \frac{1}{2} tanh(2d)
\end{aligned} \tag{7.11}$$

Bibliography

Abramowitz, M. and Stegun, I. (1965). Handbook of mathematical functions. *National Bureau of Standards, Applied Math. Series, US Govt. Printing Office, Washington DC.*

Berk, A., Zipursky, S. L., Matsudaira, P., David, B., and Darnell, J. (1999). *Molecular Cell Biology.* W H Freeman & Co (Sd).

Boden, S. A., Cavanagh, C., Cullis, B. R., Ramm, K., Greenwood, J., Jean Finnegan, E., Trevaskis, B., and Swain, S. M. (2015). Ppd-1 is a key regulator of inflorescence architecture and paired spikelet development in wheat. *Nature Plants*, 1(2):14016.

Borg, L., Smith, A., Taylor, J., and Cullis, B. (2015). *Osmotic stress experiment for Cranbrook X Halberd mapping population.*

Butler, D. G., Cullis, B. R., Gilmour, A. R., and Gogel, B. J. (2009). Mixed models for s language environments: ASReml-R reference manual. Technical Report Training Series QE02001, Department of Agriculture and Fisheries, Queensland.

Crick, F. (1958). On protein synthesis. *Symposia of the Society for Experimental Biology: The Biological Replication of Macromolecules*, (12):138–163.

Cullis, B. and Gleeson, A. (1991). Spatial analysis of field experiments-an extension to two dimensions. *Biometrics*, pages 1449–1460.

Cullis, B., Tanaka, E., Borg, L., Dolferus, R., Smith, A., and Taylor, J. (2015). Linkage map construction for the cranbrook x halberd mapping population: revisited with only snp markers. *Statistics for the Australian Grains Industry Technical Report Series.*

Cullis, B. R., Smith, A. B., and Coombes, N. E. (2006). On the design of early generation variety trials with correlated data. *Journal of Agricultural, Biological, and Environmental Statistics*, 11(4):381–393.

de los Campos, G., Naya, H., Gianola, D., Crossa, J., Legarra, A., Manfredi, E., Weigel, K., and Cotes, J. M. (2009). Predicting quantitative traits with regression models for dense molecular markers and pedigree. *Genetics*, 182(1):375–385.

Diggle, P. (2002). *Analysis of longitudinal data.* Oxford University Press, Oxford.

Gianola, D., Perez-Enciso, M., and Toro, M. A. (2003). On marker-assisted prediction of genetic value: beyond the ridge. *Genetics*, 163(1):347–365.

Gilmour, A. R., Thompson, R., and Cullis, B. R. (1995). Average information reml: an efficient algorithm for variance parameter estimation in linear mixed models. *Biometrics*, pages 1440–1450.

Green, P. and Silverman, B. W. (1994). *Nonparametric Regression and Generalized Linear Models: A roughness penalty approach (Chapman & Hall/CRC Monographs on Statistics & Applied Probability).* Chapman and Hall/CRC.

Haley, C. S. and Knott, S. A. (1992). A simple regression method for mapping quantitative trait loci in line crosses using flanking markers. *Heredity*, 69(4):315–324.

Henderson, C. R. (1950). Estimation of genetic parameters. 6(2):186–187.

Henderson, C. R. (1975). Best linear unbiased estimation and prediction under a selection model. *Biometrics*, pages 423–447.

Jansen, R. C. and Stam, P. (1994). High resolution of quantitative traits into multiple loci via interval mapping. *Genetics*, 136(4):1447–1455.

Jowett, G. H. (1952). The accuracy of systematic sampling from conveyor belts. *Applied Statistics*, pages 50–59.

Kammann, E. and Wand, M. P. (2003). Geoadditive models. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 52(1):1–18.

Knott, S. A. (2005). Regression-based quantitative trait loci mapping: robust, efficient and effective. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 360(1459):1435–1442.

Kosambi, D. D. (1944). The estimation of map distances from recombination values. *Annals of Eugenics*, 12(1):172–175.

Lander, E. S. and Botstein, D. (1989). Mapping mendelian factors underlying quantitative traits using rflp linkage maps. *Genetics*, 121(1):185–199.

Martinez, O. and Curnow, R. (1992a). Estimating the locations and the sizes of the effects of quantitative trait loci using flanking markers. *Theoretical and Applied Genetics*, 85(4):480–488.

Martinez, O. and Curnow, R. (1992b). Estimating the locations and the sizes of the effects of quantitative trait loci using flanking markers. *Theoretical and Applied Genetics*, 85(4):480–488.

Meuwissen, T. H., Hayes, B. J., and Goddard, M. E. (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics*, 157(4):1819–1829.

Ober, U., Erbe, M., Schlather, M., and Simianer, H. (2010). Kernel-based blup with genomic data.

Patterson, H. D. and Thompson, R. (1971). Recovery of inter-block information when block sizes are unequal. *Biometrika*, 58(3):545–554.

Pourahmadi, M. (1999). Joint mean-covariance models with applications to longitudinal data: Unconstrained parameterisation. *Biometrika*, 86(3):677–690.

Satagopan, J. M., Yandell, B. S., Newton, M. A., and Osborn, T. C. (1996). A bayesian approach to detect quantitative trait loci using markov chain monte carlo. *Genetics*, 144(2):805–816.

Searle, S. R. (1928). *Matrix algebra useful for statistics.* John Wiley & Sons.

Stein, M. (1999). Interpolation of spatial data: some theory for kriging.

Strandén, I. and Garrick, D. (2009). Derivation of equivalent computing algorithms for genomic predictions and reliabilities of animal merit. *Journal of dairy science*, 92(6):2971–2975.

Taylor, J. and Butler, D. (2017). R package ASMap: Efficient genetic linkage map construction and diagnosis. *Journal of Statistical Software*, 79(6).

Taylor, J. and Verbyla, A. (2011). Rpackagewgaim: Qtl analysis in bi-parental populations using linear mixed models. *Journal of Statistical Software*, 40(7).

Thompson, R. (1985). A note on restricted maximum likelihood estimation with an alternative outlier model. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 53–55.

Trow, A. (1913). Forms of reduplication:-primary and secondary. *Journal of Genetics*, 2(4):313–324.

VanRaden, P. M. (2008). Efficient methods to compute genomic predictions. *Journal of dairy science*, 91(11):4414–4423.

Verbyla, A. P., Cullis, B. R., and Thompson, R. (2007). The analysis of QTL by simultaneous use of the full linkage map. *Theoretical and Applied Genetics*, 116(1):95–111.

Verbyla, A. P., Taylor, J. D., and Verbyla, K. L. (2012). Rwgaim: an efficient high-dimensional random whole genome average (qtl) interval mapping approach. *Genetics Research*, 94(6):291–306.

Wahba, G. (1990). *Spline models for observational data.* SIAM.

Weller, J. (1987). Mapping and analysis of quantitative trait loci in lycopersicon (tomato) with the aid of genetic markers using approximate maximum likelihood methods. *Heredity*, 59(3):413–421.

Whittaker, J., Thompson, R., and Visscher, P. (1996). On the mapping of qtl by regression of phenotype on marker-type. *Heredity*, 77(1):23–32.

Xu, S. (1995). A comment on the simple regression method for interval mapping. *Genetics*, 141(4):1657.

Yang, W. and Tempelman, R. J. (2010). A bayesian antedependence model to account for linkage disequilibrum in whole genome selection. 9.

Yi, N. (2004). A unified markov chain monte carlo framework for mapping multiple quantitative trait loci. *Genetics*, 167(2):967975.

Yi, N. and Xu, S. (2002). Mapping quantitative trait loci with epistatic effects. *Genetical Research*, (79):185–198.

Zeng, Z. (1994). Precision mapping of quantitative trait loci. *Genetics*, 136(4):1457–1468.

Zhan, H. and Xu, S. (2011). Generalized linear mixed model for segregation distortion analysis. *BMC Genetics*, 12(1):97.